La dimensione temporale del parlato

Edited by: Schmid, Stephan ; Schwarzenbach, Michael ; Studer-Joho, Dieter

Atti di Congresso

AISV 2009

AISV
Associazione Italiana di Scienze della Voce

AISV 2009

5° Convegno Nazionale

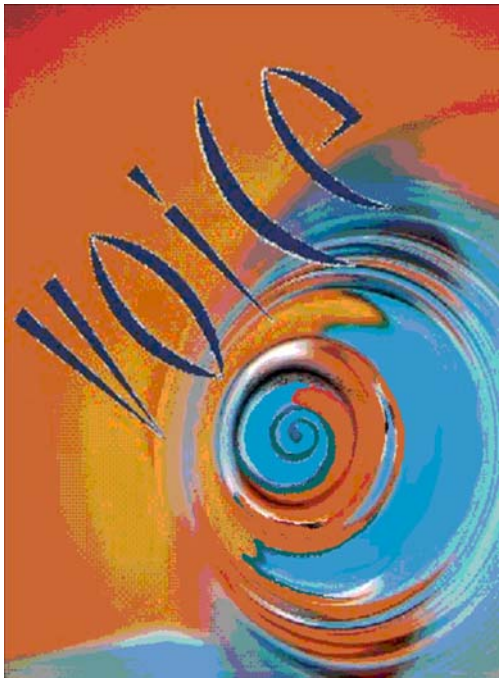
AISV - Associazione Italiana di Scienze della Voce

"LA DIMENSIONE TEMPORALE DEL PARLATO"

a cura di
Stephan Schmid
Michael Schwarzenbach
Dieter Studer

CD-rom incluso

Università di Zurigo
Kollegiengebäude
4-6 Febbraio 2009



international speech communication association

Universität Zürich
Phonetisches Laboratorium

“LA DIMENSIONE TEMPORALE DEL PARLATO”

Atti del 5° Convegno Nazionale AISV 2009

a cura di
Stephan Schmid
Michael Schwarzenbach
Dieter Studer

Università di Zurigo
Kollegiengebäude
4-6 Febbraio 2009

Universität Zürich
Phonetisches Laboratorium

“LA DIMENSIONE TEMPORALE DEL PARLATO”

Atti del 5° Convegno Nazionale AISV 2009
Università di Zurigo, *Kollegiengebäude*
4-6 Febbraio 2009

a cura di
Stephan Schmid
Michael Schwarzenbach
Dieter Studer

Via Santarcangiolese, 6
47825 Torriana (RN)

INDICE

ORGANIZZAZIONE CONVEGNO

ORGANIZZAZIONE GENERALE

Stephan Schmid

COMITATO SCIENTIFICO

Cinzia Avesani (ISTC – CNR, Padova)
Pier Marco Bertinetto (Scuola Normale Superiore, Pisa)
Silvia Calamai (Università degli Studi di Siena)
Piero Cosi (ISTC – CNR, Padova)
Francesco Cutugno (Università Federico II, Napoli)
Amedeo De Dominicis (Università della Tuscia, Viterbo)
Mauro Falcone (Fondazione Ugo Bordoni, Roma)
Barbara Gili-Fivela (Università degli Studi di Lecce)
Michele Loporcaro (Università di Zurigo)
Giovanna Marotta (Università degli Studi di Pisa)
Pietro Maturi (Università Federico II, Napoli)
Maurizio Omologo (FBK – IRST, Trento)
Andrea Paoloni (Fondazione Ugo Bordoni, Roma)
Antonio Romano (Università di Zurigo)
Luciano Romito (Università della Calabria, Arcavata di Rende)
Pier Luigi Salza (Loquendo S.p.A., Torino)
Renata Savy (Università degli Studi di Salerno)
Carlo Schirru (Università di Sassari)
Stephan Schmid (Università di Zurigo)
Mario Vayra (Università di Bologna)
Claudio Zmarich (ISTC – CNR, Padova)

SEGRETERIA DEL CONVEGNO

AISV 2009
Phonetisches Laboratorium der Universität Zürich
Rämistrasse 71
CH-8006 Zurigo
Tel. 0041 44 634 3001 - Fax 0041 44 634 6948
E-mail: aisv2009@pholab.uzh.ch
URL: http://www.pholab.uzh.ch/aisv2009.html

WEBMASTER

Stephan Schmid
Dieter Studer

ORGANIZZAZIONE LOCALE

Vincenzo Faraoni
Lorenzo Filipponio
Michele Loporcaro
Nadia Nocchi
Susanne Oberholzer
Tania Paciaroni
Dieter Studer

STAFF

Marina Albertini
Camilla Bernardasci
Francesca Beyeler
Francesco Cangemi
Sarah Heim
Fritz Herrmann
Endri Llanaj
Luana Massaro
Lucia Picuccio
Michael Schwarzenbach

SPONSOR

Hochschulstiftung der Universität Zürich
Zürcher Universitätsverein (ZUNIV)
Phonetisches Laboratorium der Universität Zürich
Phonogrammarchiv der Universität Zürich
Harman/Becker Automotive Systems Gmbh
Förderverein “Amici del Liceo Artistico”

“LA DIMENSIONE TEMPORALE DEL PARLATO”

PREMESSA

Il 5° Convegno Nazionale dell’Associazione Italiana di Scienze della Voce si è svolto, per la prima volta nella breve storia dell’AISV, all’estero, e più precisamente all’Università di Zurigo. Dal 4 al 6 febbraio si sono riuniti ricercatori italiani, svizzeri, inglesi, tedeschi e francesi per discutere del tema della ‘dimensione temporale del parlato’ nei suoi più svariati aspetti.

Com’è noto, la dimensione temporale è un elemento costitutivo della comunicazione orale, che interviene sia nella produzione che nella percezione della parola. A livello segmentale la dimensione temporale determina non solo fenomeni come durata e quantità (particolarmente interessanti in ambito italo-romanzo), ma più in generale la pianificazione e il controllo dei gesti articolatori. A livello soprasegmentale la dimensione temporale caratterizza tra l’altro l’allineamento dei contorni intonativi con le parti dell’enunciato. Inoltre, la velocità di eloquio costituisce un importante parametro in ambito forense, ma essa può essere analizzata anche da un punto di vista ‘diacronico’ (attraverso l’evoluzione del parlato dei mass media). Infine, i vari aspetti del *timing* sono di notevole rilevanza per la tecnologia del linguaggio, in particolare nel campo della sintesi della voce. Questi sono stati soltanto alcuni degli aspetti trattati nei numerosi contributi presentati durante le tre fitte giornate di lavoro, sia sotto forma di relazione orale sia come poster.

Per volontà degli organizzatori del convegno, particolare attenzione è stata rivolta alla fenomenologia del ‘ritmo’. La tavola rotonda del primo giorno su *Different ways of analyzing speech rhythm*, introdotta e moderata magistralmente dal prof. William Barry, ha permesso un confronto e uno scambio di opinioni tra alcuni specialisti in questo ambito di ricerca. Ad esempio, Volker Dellwo ha messo in evidenza come qualsiasi studio del ritmo debba tener conto anche della velocità di eloquio. Chris Lee ha invece proposto un approccio in termini di ‘prominenza’ che prende in considerazione non solo durata, ma anche F0 e intensità. Infine, Antonio Romano ha presentato alcuni studi recenti sul ritmo che sono stati condotti da riceratori italiani. I quattro contributi alla tavola rotonda aprono il presente volume.

La tematica del ritmo è stata ripresa all’ultimo giorno nella relazione plenaria di Eric Keller dal titolo *From sound to rhythm expectancy*. Lo studioso svizzero ha da un lato sviluppato alcuni argomenti discussi durante il *workshop* da lui organizzato per il congresso ICPhS di Saarbrücken nel 2007; dall’altro lato le sue riflessioni hanno testimoniato la ricca esperienza di una vita dedicata alla ricerca sulla prosodia. Della conferenza di Eric Keller si trova la presentazione .ppt nel CD-ROM.

La relazione plenaria di Martin Meyer, in apertura dei lavori scientifici del primo giorno, verteva invece sui meccanismi neuronali coinvolti nella percezione del parlato, focalizzando soprattutto l’aspetto temporale del segnale acustico. Questa impressionante revisione dello stato dell’arte nelle neuroscienze è presente negli Atti come articolo co-

firmato da Cyrill Ott e si contraddistingue anche per la ricchissima bibliografia.

Come di consueto, anche in questo convegno non si è parlato solo di un tema specifico, ma ai partecipanti è stata offerta anche una variegata rassegna delle ricerche attualmente in corso nei settori disciplinari coltivati dai membri dell'AISV – dalla linguistica, fonetica e fonologia all'ambito forense alle varie tecnologie del parlato, ecc. All'interno di questa vocazione interdisciplinare (che è parte integrante dell'identità della nostra associazione), dall'incontro zurighese sono comunque emerse due aree di ricerca che attualmente godono di un notevole interesse. La prima di queste aree di ricerca rientra in una delle tematiche da sempre presenti nei convegni AISV e si colloca all'interfaccia tra dialettologia e fonetica sperimentale, a riprova del fatto che il patrimonio linguistico dell'Italia costuisce tuttora un oggetto di ricerca di grandissima rilevanza scientifica. La seconda area di ricerca, di tradizione più recente, sembra invece nascere all'insegna della globalizzazione e ha come oggetto la percezione e l'acquisizione delle strutture sonore in una seconda lingua; dallo sviluppo che questi studi hanno sperimentato negli ultimi anni anche in Italia si evince che la loro ragione non si esaurisce nelle eventuali applicazioni pratiche, ma che questa linea di ricerca ha dei risvolti teorici considerevoli per la comprensione della facoltà umana del linguaggio.

Un convegno non è fatto solo di scienza. Le tre giornate zurighesi hanno offerto un'occasione non solo per incontrare amici di vecchia data, ma anche di fare nuove conoscenze durante gli eventi sociali – quali il rinfresco al *Romanisches Seminar*, la cena sociale oppure durante una della pause caffè nel *Lichthof*. Un momento particolarmente emozionante è stato sicuramente il bellissimo concerto del 'Trio Fontane' nell'Aula Magna.

A differenza del convegno precedente, AISV 2009 non conteneva un'apposita sezione riservata ai dottorandi. Ciononostante è stata numerosa la partecipazione sia attiva che passiva di giovani provenienti da varie Università – sicuramente un buon auspicio per il futuro delle scienze della voce.

PREMIO FRANCO FERRERO

Come nelle due edizioni precedenti, anche quest'anno è stato assegnato il 'Premio Franco Ferrero' all'autore (studente o dottorando) del miglior articolo pubblicato negli Atti del 3° e del 4° Convegno AISV. Il 'Premio Ferrero 2009' è stato consegnato durante l'apertura del convegno dal prof. Andreas Fischer, Magnifico Rettore dell'Università di Zurigo. Sono stati premiati

– nella categoria 'Linguistica, Fonetica, Fonologia':

PAOLO MAIRANO

per l'articolo "Lingue isosillabiche e isoaccentuali: misurazioni strumentali su campioni di italiano, francese, inglese e tedesco", pubblicato da Paolo Mairano & Antonio Romano negli Atti del 3° Convegno AISV (2006) tenutosi a Povo (Trento);

– nella categoria 'Tecnologie del Parlato':

GIACOMO SOMMAVILLA

per l'articolo "SMS-Festival: un nuovo ambiente di lavoro per la sintesi vocale da testo scritto", pubblicato da Giacomo Sommavilla, Carlo Drioli, Piero Cosi & Graziano Tisato negli Atti del 3° Convegno AISV (2006) tenutosi a Povo (Trento).

RINGRAZIAMENTI

Un ringraziamento particolare va ai membri del comitato scientifico per la valutazione degli Abstract e la revisione dei lavori che in questo volume sono contenuti. Grazie anche a Piero Cosi e a Luciano Romito, interlocutori dell'AISV sempre disponibili per discutere qualsiasi dettaglio organizzativo.

Mi preme ringraziare tutte le persone dell'Università di Zurigo che in un modo o l'altro hanno contribuito alla riuscita del convegno, a cominciare dal Magnifico Rettore, Prof. Andreas Fischer, per le gentili parole di benvenuto.

Ringrazio inoltre

- il dott. Maximilian Jaeger, delegato del Rettore, per la sua disponibilità riguardo all'organizzazione logistica,
- il personale del *Veranstaltungsdienst Zentrum* per l'assistenza tecnica,
- i musicisti del 'Trio Fontane',
- la dott.ssa Katharina Maier-Troxler e Vera Ziswiler per l'aiuto nell'organizzazione del rinfresco al *Romanisches Seminar,*
- tutti gli studenti e assistenti di linguistica italiana per le ore di servizio passate nella segreteria del convegno,
- gli Sponsor che con il loro contributo finanziario hanno agevolato la realizzazione delle tre giornate zurighesi.

Infine, ringrazio di cuore Michele Loporcaro per aver sostenuto dall'inizio alla fine l'avventura del convegno zurighese – nonché Didi e Michi, cari compagni di lavoro.

Stephan Schmid

TAVOLA ROTONDA:

“DIFFERENT WAYS OF ANALYZING SPEECH RHYTHM”

RHYTHM MEASURES IN RETROSPECT. REFLECTIONS ON THE NATURE OF SPOKEN-LANGUAGE RHYTHM

William J. Barry
Institut für Phonetik, Universität des Saarlandes
wbarry@CoLi.Uni-SB.DE

1. ABSTRACT

An account is offered of the background and developments leading to the present state of research into quantitative rhythm analysis. The two most influential rhythm metrics of the past decade are explained, what they achieve and fail to achieve are discussed, and modified metrics of the same ilk are described. Finally, the needs for future rhythm research are considered.

2. INTRODUCTION

The 'rhythm of spoken language' appears to be a concept that is accepted as a given fact that needs no further definition. Yet even the most superficial discussion among a group of people who agree that there *is* such a thing normally leads to a wide range of opinions on *what* it is. The divergence in opinion seems to stem from the particular focus that is adopted; whether rhythm is understood to be the regular repetition of a particular *pattern* in the sense of poetic metre, or the regular *strong beat* irrespective of how many weaker beats form the pattern between them; or the focus may be on the way in which *any individual utterance* is realized, such that the rhythm of an utterance is seen as the way the sequence of words is spoken by a particular person in a particular context. Such different views can result in (i) *none* of a number of realizations of a particular sequence of words being considered rhythmical (because they do not conform to a metrical pattern), (ii) *some* being considered more or less rhythmical than others (because some were produced with very strong regular beats of accented syllables while others were produced with less strongly accented and/or less regularly accented syllables), or (iii) *all* being considered rhythmical in their own particular way. Probably the only consensus to be achieved is that the 'rhythm of spoken language' is something that speakers *produce* and listeners *perceive*.

The *production* aspect is presumably the foundation upon which instrumental phoneticians base their assumptions that they would be able to find something in the speech signal to measure and use as evidence. Certainly the *perception* aspect underlies most phoneticians' experience of speech rhythm; Lloyd James' (1940) observation that his *rhythmic impressions* of spoken French and English differed fundamentally – namely a 'machine gun' vs. a 'morse code' impression – is probably the most common point of departure for anyone writing about rhythm. Interestingly, Lloyd James was not claiming that his auditory impression provided a basis for acoustic analysis, nor was he making a claim for a universal rhythmic typology. That generalization is usually attributed to Kenneth Pike (1946), who suggested a dichotomy of 'syllable-timed' and 'stress-timed' rhythm types. Peter Ladefoged (1975) took up Bernard Bloch's (1950) view of Japanese timing as being determined by the mora structure and added 'mora-timing' to the ostensibly all-encompassing rhythm universals.

It is very doubtful that the originators of the rhythm-typology concept envisaged an isochronous underlying timing mechanism onto which the foot-, syllable- or mora-producing part of our articulation locked, However, the inevitable simplification of ideas that results from practical application and the operationalisation of analysis methods meant that syllable-/stress-*timing* became inextricably associated with syllable-/stress-*isochrony*. The language-teaching orientation of the 'British School' meant that exercises in approximating isochrony of stressed syllables (with the concomitant compression and reduction of intervening unstressed syllables) helped to release L2 learners of English from their rhythmically incorrect, 'non-reducing' L1 production habits. Crucially for instrumental phonetics, the isochrony concept provided a criterion for *acoustic* analysis, and the number of studies that have pursued isochrony in so many different ways[1] is testimony to its plausibility. Any non-naive search for isochrony will, of course, take into account both the intrinsic durational differences of segments that comprise syllables and the complexity of the syllable structure. Thus, the search was always statistical, a search for a stronger *tendency* towards equal syllable durations or equal stress-group duration. However, the search was unsuccessful, and the most positive conclusion that can be drawn is that the non-results provided a fruitful ground for the intellectual activity of probing the complexities of the timing of speech events. The many rationalisations tell a fascinating story, most cogently and edifyingly summarized and discussed by Bertinetto (1989).

It is instructive to briefly consider where any underlying isochrony, if it should exist, disappears to: the type and number of sound segments in a syllable, the number and complexity of syllables in a word, lexical prosodic effects on syllables and phrase-based prosodic effects on words result in ubiquitous variability and a wide range of durations for whatever linguistic unit is measured. However, as is apparent from discussion in Bertinetto (1981), Dasher & Bolinger (1982) and Dauer (1983), the variability effects differ in a non-random manner from language to language. For example, purely in the temporal domain, which impinges directly on 'isochrony' measures, there are languages that have a wider range of syllable structures than others (i.e. more complex ones as well as the simpler ones). There are those that have a larger proportion of polysyllabic words in their lexicon vs. those that favour mono- or disyllables. There is the well-known tendency in some languages for lexically unstressed or phrasally unaccented syllables to be strongly reduced (in both quality and quantity), whereas others maintain a stronger resistance to either sort of reduction. Some languages show a reluctance to de-accent words conveying 'given' information while others de-accent (and consequently reduce) more readily. These different properties can co-occur in languages, resulting in a "conspiracy of factors" (Bertinetto, 1989) which push them towards one end or the other of the variability continuum, supporting the old 'syllable-timed' vs. 'stress-timed' dichotomy. However, a language might be more variable in one factor and less variable in another, giving a mixed result.

[1] Shen & Peterson, 1962; Bolinger, 1965; Delattre, 1966; Lehiste, 1977; Faure, Hirst & Chafcouloff, 1980; Pointon, 1980; Nakatani, O'Connor & Aston, 1981; Wenk & Wioland, 1982; Roach, 1982; Dauer, 1983; Hoequist, 1983; Manrique & Signorini, 1983; Dauer, 1987; Eriksson, 1991. These studies document the wide and continual interest, though the list is by no means exhaustive.

This fits well with the fact that not all linguists agree on the allocation of languages to the traditional rhythm types.

The logical corollary of this language-dependent breakdown of factors influencing variability is a consideration of rhythm typology from the *variability* as opposed to the *isochrony* perspective. This was precisely what happened, though it was more than a decade after the intellectual groundwork discussed above when the first quantitative studies appeared. Independently of one another, researchers in Paris (cf. Ramus *et al.*, 1999) and Cambridge (Low *et al.*, 2000; Grabe & Low, 2002) applied variability metrics to a number of 'rhythmically' different languages and confirmed that they could be separated in a way which corresponded plausibly with their assumed rhythmic type. It is these metrics and their offshoots which the next section addresses.

3. RHYTHM MEASURES BASED ON STRUCTURAL VARIABILITY

Although the mathematics differed, the rationale behind both Ramus' and Low *et al.*'s metric was the same: 1) If languages differ in the range of complexity of their syllable structures, the duration of syllables in languages with only *simple* syllable structure (traditionally the 'syllable-timed' languages) will *vary little* while in languages that also allow *complex* syllables (traditionally more likely to belong to the stress-timed category) it will vary more. 2) Consonants and vowels contribute to syllabic variability in different ways and to a potentially differing degree. Therefore there should be separate measures for C and V intervals. This separation conveniently provided two axes of variation for a graphic delimitation of the 'rhythmic space' (see figure 1).

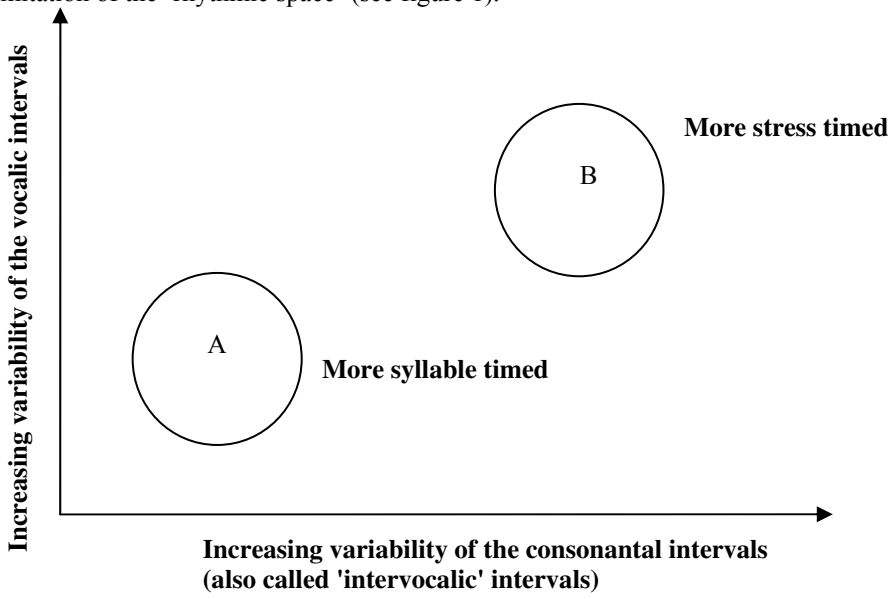


Figure 1: Separation of traditional rhythm types on the basis of measured variability of vocalic and consonantal intervals.

Ramus' variability measure was a simple standard deviation of the consonantal (ΔC) and vocalic intervals (ΔV).

E.g. *Nessuno sa come rimediare al misfatto*.

/nɛs'sunɔ sa kɔmɛ rime'djarɛ al mis'fattɔ/

1 2 3 4 5 6 7 8 9 10 11 12 13 intervals for ΔC.

/nɛs'sunɔ sa kɔmɛ rime'djarɛ al mis'fattɔ/

and 1 2 3 4 5 6 7 8 9 10 11 12 13 intervals for ΔV.

He added a third measure which was unrelated to variability but which also reflected the relative complexity of the syllable structure, namely %V, the summed vowel duration relative to the total duration of the utterance.

In contrast to Ramus' simple measure of global variability, Ee Low's measure, which was named 'Pairwise Variability Index (PVI), had a sequentiality component. The durational differences from one interval to the next were averaged across the utterance. This is expressed in the formula:

$$PVI = \left[\sum_{k=1}^{m-1} \left| dk - dk + 1 \right| / (m-1) \right]$$

For the consonantal intervals PVI-C this would be calculated as follows:

Nessuno sa come rimediare al misfatto.

/nɛs'sunɔ sa kɔmɛ rime'djarɛ al mis'fattɔ/

1−2 3−4 5−6 7−8 9−10 11−12

2− 3 4−5 6− 7 8−9 10−11 12−13

The vocalic intervals (PVI-V) would be calculated as:

Nessuno sa come rimediare al misfatto

/nɛs'sunɔ sa kɔmɛ rime'djarɛ al mis'fattɔ/

1−2 3−4 5−6 7−8 9−10 11−12

2−3 4−5 6−7 8−9 10 −11 12−13

However, to correct for possible tempo changes within an utterance, a normalized PVI-V was defined:

$$PVI = 100 \times \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| / (m-1) \right]$$

4. DO THE MEASURES SEPARATE LANGUAGES?

As stated at the end of section 1, both Ramus *et al.* (1999) and Grabe and Low (2002) presented data which illustrated that their measures supported groupings of languages that had traditionally been allocated to the same rhythm type. The results also supported the idea that rhythm types were not discrete categories but that languages were located within a continuous rhythm space, defined by the consonantal and vocalic axes. However, a brief comparison of the figures 2 and 3 shows that, depending on the metric used, the same language can occupy different positions relative to other languages. Comparing the order of the languages along the consonantal variability axis, which is used in both diagrams, we see that the order of languages with the Grabe/Low PVI-C measure (compare fig. 2) is:

Spanish < Japanese < French < Catalan < Polish < English < German < Dutch.

Ramus' ΔC measure orders the same languages (compare fig. 3):

French < German < Dutch = Spanish < Japanese < English < Catalan < Polish

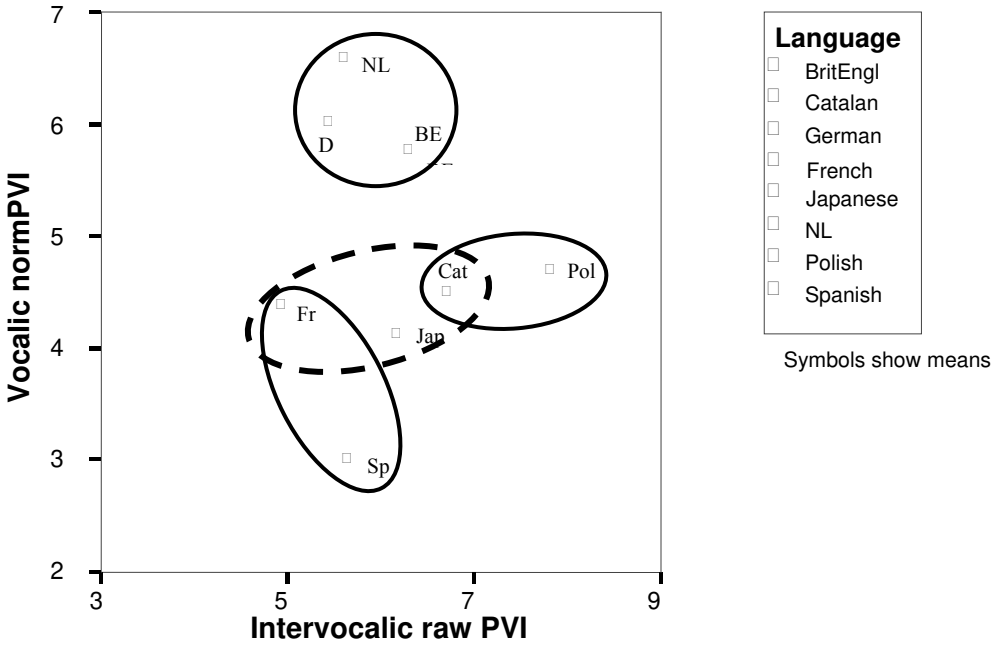


Figure 2: PVI-values for selected languages (after Grabe & Low, 2002)

Inspired by these two publications, studies quickly followed which uncovered the dependency of the 'rhythm measures' for any one language on the material used, the speakers selected and the speaking style elicited (cf. Barry *et al.*, 2003). New measures have been derived to correct for perceived weaknesses in the original ones. Since variation increases with the average duration of the units measured, relating the two was an important step taken by Dellwo (2006) who introduced the coefficient of variation (Varco) – the standard deviation divided by the mean – into rhythm analysis. This step alone, though it

counteracts consistent speech-rate differences, cannot correct for other problems. Bertinetto & Bertini (2008) modified the PVI measure to normalize for the number of segments constituting the vocalic and consonantal intervals, arguing that languages differ in the degree to which they 'control' the segmental production in a syllable. The results within this Control vs. Compensation (CC) theory are clearly similar though not identical to the traditional division into syllable- and stress-timing. Since normalisation brings a greater likelihood of a reduced value for the consonantal axis in languages where the syllable-complexity range is greater, traditional stress-timed languages have a low C-value to V-value ratio and are called 'compensating' languages. Traditional syllable-timed languages have a more equal ratio and are called 'controlling languages'.

Part of the rationale for the original PVI and Delta measures was the different contribution of the consonantal and vocalic parts of a syllable to durational variability. One theoretical flaw in these separate measures is the separation of each consonantal interval from the vowel to which the consonants syllabically belong. Since the syllable is generally seen as the basic rhythmic unit, either by itself or as part of a stress group, the division of the syllable would seem to destroy 'rhythmic information'. It is not surprising, therefore, that the syllable has been used as the unit for many metrics, and that some rhythm researchers have gone against the priciple of consonant and vowel separation to apply the PVI measure fruitfully at syllabic level (Deterding, 2001) and even at both syllable and foot-level for a two-dimensional analysis with what are traditionally seen as opposing rhythmic building blocks (Asu & Nolan, 2006; Nolan, 2009).

The most comprehensive critique of rhythm measures to date is probably that presented by Arvaniti (2009), who combines a discussion of points raised previously by others with fresh points of view derived from her own observations and analyses. One important aspect that she addresses which has been rather neglected in many studies is the statistical basis of differences between languages and groupings of languages. Given the variation in measures due to the speech material analysed, the selection of speakers and the speaking styles, the average position in the 'rhythm space' of an analysis sample for one language and the distance to the position of a sample from another language may well be determined by chance rather than typological difference. Arvaniti's own statistical analysis of data available to her revealed rather unreliable results.

In a paper (unpublished in the form presented) at the University College London *Workshop on Empirical Approaches to Speech Rhythm* in March 2008, she warned of the danger of interpreting graphs in terms of traditional typological categories. Figures 2 and 3 illustrate this danger. In figure 2 the solid line ellipses grouping Spanish with French, Polish with

Catalan and Dutch with English and German appear to support the syllable-timed vs. stress-timed distinction, with further support for a 'mixed' category. Japanese, the traditionally mora-timed language is left outside the groupings. The dashed-line ellipse, on the other hand, draws attention to the fact that German and Catalan are both closer to Japanese than they are to French and Polish, respectively, with which they had originally been grouped. The two solid-lined ellipses in figure 3, which shows the Ramus values for the same languages and utterances as shown in figure 2, once again suggest support for grouping the traditional syllable-timed languages together and the stress-timed languages together. This time, Polish and Catalan cannot be placed in one group. Japanese is also ungrouped again, although it appears to be closer to French than German is to Dutch. The

dashed-line ellipses, on the other hand, highlight the fact that German is as close, if not closer to Catalan than to Dutch and English.

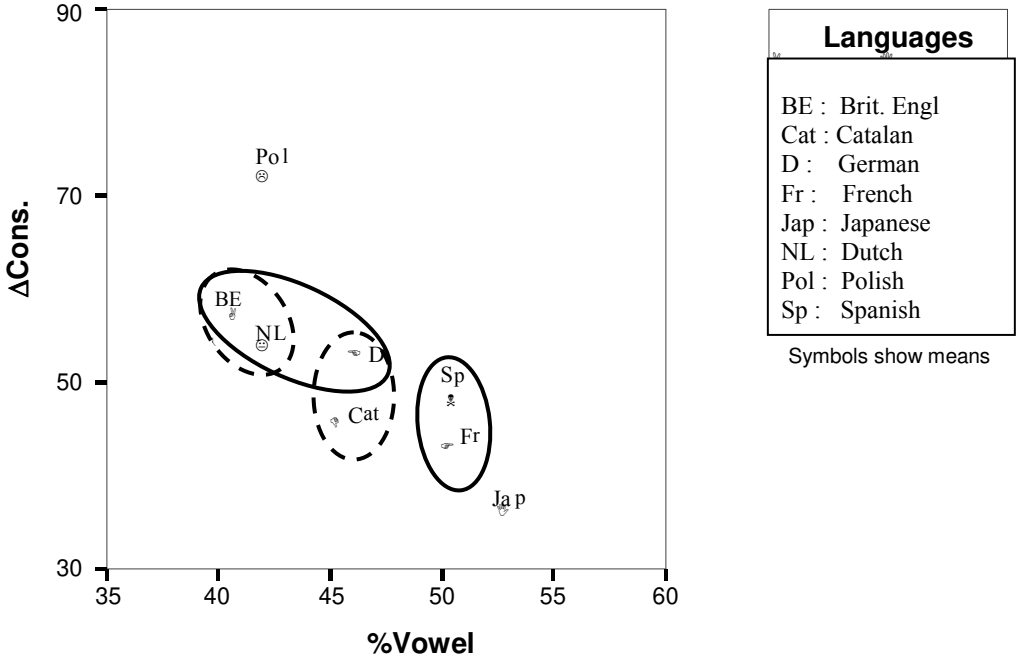


Figure 3: 'Ramus values' for selected languages (after Grabe & Low, 2002)

5. ARE WE REALLY MEASURING RHYTHM?

It is almost impossible to give a yes or no answer to such a question. We started the introductory discussion with a statement that spoken language rhythm is open to a number of definitions. What we *can* say, is that the measures all capture systematic structural sound differences between languages, as long as the material measured is representative of each of the languages, and the speakers are fluent, representative speakers of that language.

PVI-V captures the degree to which *consecutive vowel durations* vary; i.e., it is sensitive to the realisation of phonemically long vs short vowels, to lexical stress and phrasal accentuation effects, and even to phonetic variation due to differences in degree of opening. PVI-C captures the degree to which *consecutive intervocalic interval durations* vary (i.e., single consonants or clusters), and to deletions in cases of reduction. Inasmuch as the average measure for an utterance is the result of *sequential pair differences*, the measure is sensitive to short-long differences, which are part of what makes something rhythmic. In contrast, the ΔV and ΔC measures are insensitive to any adjacency effects; they are bereft of any link to rhythmic structures. Due to the separate calculation of vocalic and intervocalic variability, none of the measures are sensitive to adjacent changes in syllable or foot units (the building blocks of rhythmic structures), unless the PVI metric is applied explicitly to syllables and feet (cf. Deterding, 2001; Asu & Nolan, 2006; Nolan & Asu,

2009). On the other hand, analyses of different poetic metres by Barry *et al.* (2009) *did* show an interpretable sensitivity to the difference between iambic or trochaic metres on the one hand and dactylic or anapest (or more complex) metres on the other. This was presumably due to the systematic alternation of strengthened and weakened vocalic *and* consonantal intervals *together,* producing a quasi syllabic effect.

One important aspect of all the rhythm measures so far developed is their total dependence on *durational* properties. Thus, if rhythm is to be captured by any of them it must depend completely on durational properties of the utterance. Without trying to pre-decide what spoken-language rhythm really is, we can probably achieve a broad consensus about its dependence on "patterns of more prominent and less prominent syllables". Thus we are bound also to accept that rhythmic impressions are dependent to a greater or lesser degree on *all* the signal properties that contribute to prominence. Since there is long established experimental evidence that all four basic acoustic attributes of the speech signal – duration, F0, intensity and spectral shape – can influence a listener's impression of relative prominence, both in lexical stress and phrasal accent judgments, we must assume that they also contribute to the impression of rhythmicity. There is increasing awareness that measuring durations alone must distort the picture of speech or language rhythm that is created (Arvaniti, 2009; Kohler, 2009; Nolan & Asu, 2009). Barry *et al.* (2009) showed that listeners' judgment of *degree* of rhythmicity (of poetic metre) was strongly influenced by F0 as well as duration. The influence of intensity and vowel quality appeared, however, within the experimental paradigm used, to be very weak. Niebuhr (2009) has also demonstrated a rhythmically interpretable dependence on F0 in the judgment of syllabic prominence.

It is, of course, not necessarily the case that impressions of rhythm in poetic metre are subject to the same rules as the 'rhythm' of prose. However, with information-structuring accents heavily reliant on their tonal shape both for their linguistic identity and for their communicative function, tonally transmitted prominence appears an essential part of the normal prominence patterning of spoken language.

6. SOME FINAL THOUGHTS

One thing is certain: the rhythm of spontaneous speech is, as a rule, not comparable to the rhythm of consciously structured poetry. However, rhetorically skilled speakers can produce phrases that, if extracted from the speech, could well serve as a line in a poem. I would therefore concur with Kohler (2009) that not all speech that one might record and analyse can be regarded as an instance of 'speech rhythm'. Similarly, given the limitless choice of lexicon-grammar combinations that make up the utterances of any one language, not all the possible utterances can be considered equally typical of the rhythmic patterning of that language (whatever 'rhythm' actually turns out to be). If the 'rhythm' of a spoken language is to be quantitatively captured, so that the rhythm typology question can be satisfactorily pursued, there has to be a very large number of fluently produced utterances, produced by informants who are accepted as skilled and representative speakers. Crucially too, there has to be a metric which takes not only duration into account but also includes at least F0 in a complex measure of fluctuating prominence (cf. Lee & Todd, 2004 for an auditory model based process for deriving 'rhythmograms').

Finally, we need to bear the listeners in mind, without whom production patterns make no sense. Recent work by Wagner (2008) and observations by Arvaniti (2009) have brought

the idea of *perceptual grouping* as the foundation stone of rhythmic structure into the discussion. The prosodic means for triggering the grouping process can differ from language to language. Thus, understanding the prosodic structures of a language rather than number-crunching with a 'rhythm metric' would be the key to identifying its rhythmic type.

ACKNOWLEDGEMENTS

The research underlying this discussion paper was supported by the German Research Council (Deutsche Forschungsgemeinschaft) grant BA 737/9-7.

7. REFERENCES

Asu, E.L. & Nolan, F. (2006), Estonian and English rhythm: a two-dimensional quantification based on syllable and feet, in *Speech Prosody 2006*, Dresden: TUDpress, 249-252.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm, *Phonetica*, 66, 46-63.

Barry, W. J., Andreeva, B., Russo, M., Dimitrova, S. & Kostadinova, T. (2003), Do rhythm measures tell us anything about language type?, in *Proceedings of the 15th International Congress of Phonetic Sciences,* Barcelona, 2693-2696.

Bertinetto (1981), *Strutture prosodiche dell'italiano*, Firenze: Accademia della Crusca.

Bertinetto, P.M. (1989), Reflections on the dichotomy 'stress' vs. 'syllable-timing', *Revue de Phonétique Appliquée,* 91-93, 99-130.

Bertinetto, P.M. & Bertini, C. (2008), On modeling the rhythm of natural languages, in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, May 6-9, 2008, 427-430.

Bloch, B. (1950), Studies in colloquial Japanese IV: Phonemics, *Language*, 26, 86-125.

Bolinger, D.L. (1965), *Forms of English: Accent, Morpheme, Order*, Cambridge, Massachusetts: Harvard University Press.

Dasher, R. & Bolinger, D. (1982), On pre-accentual lengthening, *Journal of the International Phonetic Association*, 12, 58-69.

Dauer, R.M. (1983), Stress-timing and syllable-timing re-analysed, *Journal of Phonetics,* 11, 51-62.

Dauer, R. (1987), Phonetic and phonological components of language rhythm, in *Proceedings of the 11th International Congress of Phonetic Sciences*, Talinn, Estonia, 447-450.

Delattre, P. (1966), A comparison of syllable-length conditioning among languages, *International Review of Applied Linguistics*, 4, 183-198.

Dellwo, V. (2006), Rhythm and speech rate: a variation coefficient for DeltaC, in *Language and language processing: proceedings of the 38th Linguistics Colloquium* (P. Karnowski & I. Szigeti, editors*)*, Frankfurt: Lang, 231-241.

Deterding, D. (2001), The measurement of rhythm: a comparison of Singapore and British English, *Journal of Phonetics*, 29, 217-230.

Eriksson, A. (1991), *Aspects of Swedish speech rhythm*, Gothenburg Monographs in Linguistics, 9, University of Göteborg.

Faure, G. Hirst, D.J. & Chafcouloff, M. (1980), Rhythm in English: Isochronism, pitch, and perceived stress, in *The Melody of Language* (L.R. Waugh & C.H. van Schooneveld, editors), Baltimore: University Park Press, 71-79.

Grabe, E. & Low, E.L. (2002), Durational Variability in Speech and the Rhythm Class Hypothesis, in *Papers in Laboratory Phonology 7* (C. Gussenhoven & N. Warner, editors), Berlin: Mouton de Gruyter, 515-546.

Hoequist, C.J. (1983a), Durational correlates of linguistic rhythm categories, *Phonetica*, 40, 19-31.

Hoequist, C.J. (1983b), Syllable duration in stress-, syllable- and mora-timed languages, *Phonetica*, 40, 203-237.

Kohler, K. (2009), Rhythm in speech and language. A new research paradigm, *Phonetica*, 66, 46-63.

Ladefoged, P. (1975), *A Course in Phonetics*, New York: Harcourt Brace Jovanovich.

Lee, C.S. & McAngus Todd, N.P. (2004), Towards an auditory account of speech rhythm: application of a model of the auditory 'primal sketch' to two multi-language corpora, *Cognition*, 93, 225-254.

Lehiste, I. (1977), Isochrony reconsidered, *Journal of Phonetics*, 5, 253-263.

Lloyd James, A. (1940), *Speech signals in telephony*, London: Sir I. Pitman & Sons.

Low, E.L., Grabe, E. & Nolan, F. (2000), Quantitative characterisations of speech rhythm: 'syllable-timing' in Singapore English, *Language and Speech*, 43, 377-401.

Manrique, A.M.B. & Signorini, A. (1983), Segmental reduction in Spanish, *Journal of Phonetics*, 11, 117-128.

Nakatani, L.H., O'Connor, J.D. & Aston, C.H. (1981), Prosodic aspects of American English speech rhythm, *Phonetica*, 38, 84-105.

Niebuhr, O. (2009), Fundamental frequency-based rhythm effects on the perception of local syllable prominence, *Phonetica*, 66, 95-112.

Nolan, F. & Asu, E.L. (2009), The pairwise variability index and coexisting rhythms in language, *Phonetica*, 66, 64-77.

Pike, K. (1946), *The Intonation of American English*. 2nd edition, Ann Arbor: University of Michigan Press.

Pointon, G.E. (1980), Is Spanish really syllable-timed? *Journal of Phonetics*, 8, 293-304.

Ramus, F., Nespor, M. & Mehler, J. (1999), Correlates of linguistic rhythm in the speech signal, *Cognition*, 73, 265-292.

Roach, P. (1982), On the distinction between 'stress-timed' and 'syllable-timed' languages, in *Linguistic Controversies* (D. Crystal, editor), London: Arnold, 73-79.

Shen, Y. & Peterson, G.G. (1962), Isochronism in English, *University of Buffalo Studies in Linguistics, Occasional Papers*, 9, 1-36.

Wenk, B. & Wioland, F. (1982), Is French really syllable-timed?, *Journal of Phonetics*, 10, 193-216.

CHOOSING THE RIGHT RATE NORMALIZATION METHOD FOR MEASUREMENTS OF SPEECH RHYTHM

Volker Dellwo
Division of Psychology and Language Sciences, University College London
v.dellwo@ucl.ac.uk

1. ABSTRACT

Some acoustic correlates of language rhythm are durational characteristics of consonants and vowels. The present study investigates the influence of speech rate on these acoustic correlates. In experiment I four widely applied correlates of speech rhythm (%V, ΔC, nPVI and rPVI) were correlated with the rate of consonantal and vocalic intervals using speech from five different languages (Czech, English, French, German, Italian) that was characterized by high tempo variability within each language (very slow to very fast produced speech). It was found that rhythm measures based on consonantal interval durations (ΔC, rPVI) correlate negatively with rate measures and that rhythm measures based on vocalic intervals (%V, nPVI) are not influenced by rate. In experiment II the effectiveness of rate normalization procedures on the rate dependent measures, ΔC and rPVI, was tested by correlating these measures with speech rate before and after normalization using the same speech data as in Experiment 1. ΔC was normalized by logarithmically transforming the consonantal interval durations and rPVI was normalized by previously proposed ways for the normalization of nPVI. It was found that rate effects on ΔC and rPVI could be normalized for effectively using the suggested rate normalization procedures. In Experiment III it was tested whether rate normalized measures of speech rhythm support the impression that some languages can be categorized according to their auditory rhythmic characteristics (e.g. stress- and syllable-timing). Strong support for this was only found for the rate normalized rPVI why the normalized ΔC revealed mixed results. It was concluded that ΔC is less appropriate for rhythmic measurements that aim to separate languages of different rhythmic classes.

2. INTRODUCTION

The systematic study of speech rhythm began towards the beginning of the past century. It was motivated by assumptions that rhythm plays an important role in acquiring a correct pronunciation in a foreign language and thus enhancing non native speech intelligibility (James, 1929) or simply for phonetic classification purposes (Classe, 1939; Pike, 1945; Abercrombie, 1967). More recent findings demonstrating that knowledge about the rhythm of a language is crucial for predicting word boundaries (Cutler, 1997; Cutler & Norris, 1988; Kim, Davis & Cutler, 2008) or that rhythmic cues can help infants segregating between different languages when growing up in a bilingual environment (Nazi *et al.*, 1988; Ramus *et al.*, 1999) gave rise to a growing interest in the field of speech rhythm.

Rhythmic variability in speech is manifold. Languages, for example, can posses specific auditory rhythmic characteristics (James, 1929; Classe, 1939; Pike, 1945; Abercrombie, 1967; Ramus *et al.*, 1999; Grabe & Low, 2002) but there is also rhythmic variability within a language. Native-speakers can sound rhythmically different from non-

native speakers (White & Mattys, 2007; Mok & Dellwo, 2008) and different language varieties may be characterized by different rhythmic features (e.g., Singaporean and Standard Southern British English: Low, Grabe, & Nolan, 2000; Deterding, 2001). Even speakers of the same language variety may differ in speech rhythm (Dellwo & Koreman, 2008) and rhythm may vary within the same speaker depending, for example, on emotional state (Cahn, 1990). One of the most central questions in the field of speech rhythm has been how such auditory rhythmic variability can be measured in the speech signal or in other words: What are the acoustic correlates of speech rhythm? This question turned out to be difficult to answer since unlike other perceptual prosodic phenomena like intonation, which is mainly encoded by fundamental frequency variability, it seems less clear which acoustic phenomenon is responsible for the percept of speech rhythm. It is also likely that rhythm is encoded by a number of different acoustic parameters like fundamental frequency, amplitude, and duration, and possibly our perception of rhythm results from a complex interaction between those parameters. It is further unclear whether the perception of rhythm by listener groups with varying linguistic background (e.g. different native languages) is based on the same acoustic parameters in the same way. Given that listeners of different languages, for example, make different use of prosodic stress correlates (Wang, 2008) it seems conceivable that a similar situation is true in the case of speech rhythm.

Of all the rhythmic variability in speech, the variability of rhythm between languages has, without doubt, been studied most. And it is probably for this reason that measures of speech rhythm have mostly been developed to capture between-language rhythmic variability in the speech signal. How can this between-language variability be characterized? There have been numerous attempts to categorize languages that share auditory rhythmic features (Classe, 1939; James, 1929; Pike, 1945; Abercrombie, 1967; Ramus *et al.*, 1999; Grabe & Low, 2002). Such rhythmic features were metaphorically described as sounding either more like a 'machine-gun' (e.g. French, Spanish and Yoruba) or more like a 'Morse code' (e.g. English, German, and Arabic). This comparison, introduced by James (1929) and still widely used in present times, reveals the idea that some languages appear more regularly timed than others (machine-gun vs. Morse-code respectively). This regular timing was initially assumed to be manifested in regular (or quasi-isochronous) syllabic durations in machine-gun languages and irregular (or non-isochronous) syllabic durations in Morse-code languages. Languages that were assumed to have regularly timed syllables were therefore referred to as 'syllable-timed'. The assumed lack of durational syllable regularity in Morse-Code languages led to the idea that the intervals between stressed syllables (inter-stress intervals) in such languages are timed regularly (irrespective of the number of unstressed syllables they contain). Languages revealing these assumed acoustic characteristics were therefore called 'stress-timed' languages. However, it remains unclear why stress-timed languages were (and often still are) assumed to have regularly timed inter-stress intervals. No study reports auditory support for such an assumption. It thus seems conceivable that the idea of quasi-isochronously timed inter-stress intervals in Morse-code languages was created merely as an acoustic analogy to the isochronous timing of syllables in syllable-timed languages. The early assumptions that language characteristic rhythm stands in relation with the timing of syllables or inter-stress intervals is probably the reason for the fact that most of the attempts to measure speech rhythm in the acoustic signal are based on measuring

segmental durational characteristics of speech and widely disregard other durational (e.g. the timing of fundamental frequency contours) or spectral parameters (e.g. fundamental frequency variability, dynamic variability, etc.; see Tilsen and Johnson, 2008, for an alternative approach).

It seems not surprising that the earliest attempts to measure stress- and syllable-timing in the speech signal were based on measuring the durational variability of syllables and inter-stress intervals in these languages, assuming that the variability of syllable durations should be lower and the variability of inter-stress interval durations should be higher in syllable- than in stress-timed languages. Countless approaches have been carried out from the 1960s to the end of the 1980s to find evidence for this assumption, however, no support has ever been found (see Ramus *et al.*, 1999; Grabe & Low, 2002, for reviews of the literature). It thus seems that the use of the terminology 'stress-timing' and 'syllable-timing' should be discontinued and be replaced by terminology closer reflecting the auditory impression of rhythm like 'regular' vs. 'irregular' rhythm. (The present article makes a first approach to do this. However, given the continuous wide usage of the terminology 'stress' and 'syllable' timing it will here continued to be used in parallel.)

In search for acoustic regularity and irregularity in syllable- and stress-timed languages respectively, a number of studies started reporting first success when shifting the unit of analysis from syllable and inter-stress interval durations to consonantal (C) and vocalic (V) interval durations[1] (Ramus *et al.*, 1999; Grabe & Low, 2002; Dellwo, 2006; Mattys and White, 2007). It was assumed that the durational characteristics of C and V intervals are influenced by language specific phonological features and that languages sharing such features appear rhythmically similar (Bolinger, 1981; Roach, 1982; Dauer, 1983, 1987; Ramus et al, 1999; Grabe & Low, 2002). Stress-timed languages typically share the feature of a high complexity of C intervals (i.e. allowing multiple consonants in a C interval) which leads to a high variability of C interval durations. C intervals in syllable-timed languages typically consist of only one consonant. Further, stress-timed languages typically share the feature of reducing vowels to schwa, which is believed to introduce high durational variability of V intervals. Such reduction phenomena are untypical in syllable-timed languages. Given these assumptions four measures were proposed which have been widely applied in the field of speech rhythm measurements. Two measures were proposed by Ramus *et al.* (1999). They are the standard deviation of C intervals (ΔC) and the percentage over which speech is vocalic (%V). Two further measures have been proposed by Grabe & Low (2002) which calculate the average differences between consecutive C intervals and V intervals and are known as the Pairwise Variability Index (PVI). The PVI applied to V intervals was rate normalized (nPVI, discussion follows) and the PVI for C intervals was not rate normalized (thus referred to as 'raw'; rPVI). All before mentioned measures are basically influenced by the durational variability of C and

[1] A C interval is the duration of a string of consonants between two vowels (or any combination of vowel and pause) and, likewise, a V interval is a string of vowels between two consonants (or any combination of consonant and pause) in speech. Both C and V intervals may contain a syllable, word or even sentence boundary. Mind that some studies refer to C intervals as 'inter-vocalic-intervals' (Grabe & Low, 2002), however, draw no methodological difference.

V intervals: high variability leads to high values, low variability leads to low values.[2] Numerous studies have demonstrated that such variability measurements support the distinction at least of some languages into rhythmic categories (Ramus *et al.*, 1999; Grabe & Low, 2002; Dellwo & Wagner, 2003; White & Mattys, 2007; Mok & Dellwo, 2008). It thus seems conceivable that the auditory impression of rhythmic regularity and irregularity in syllable- and stress-timed languages respectively, is the result of acoustic parameters like ΔC, %V, or the PVI measures. Support for this theory has been found for the parameters ΔC and %V (Ramus *et al.*, 1999).

All measures mentioned above are based on durational characteristics of C and V intervals. Consequently speech produced at higher rates will shorten such intervals and speech at lower rates will lengthen them. Since the techniques involved in the production of speech vary widely across different sounds, it must be assumed that rate does not affect the duration of different segment types in the same way. Thus the durational characteristics like ΔC or PVI, for example, may vary when speakers speak faster or slower and such changes are unlikely to be of a linear nature (Ramus, 2003; Grabe & Low, 2002; Dellwo & Wagner, 2003; Barry *et al.*, 2003; Dellwo, 2006). Further, speech rate can influence such measures on two levels: First, given that the overall durations of C and V intervals change when speech is uttered faster or slower, it may have an influence on the within-language variability of C and V intervals. Second, the rate of C and V intervals can vary significantly between languages (Dellwo, 2008) for the same reasons that rhythmic properties vary. After all, less complex C intervals should on average be shorter thus should be produced at higher rates. For the rhythm measures this means that, given they interact with rate systematically, it may inevitably lead to situations in which an obtained acoustic rhythmic difference between two groups under observation is actually a rate difference between these groups.

A number of studies contain evidence that rhythmic measurements are influenced by rate parameters (Grabe & Low, 2002; Dellwo & Wagner, 2003; Barry *et al.*, 2003; Ramus, 2002; Dellwo, 2006; White & Mattys, 2007; Dellwo, 2008). These studies, however, have shortfalls: They are either based on data with a very poor rate variability (Grabe & Low, 2002; White & Mattys, 2007; Ramus, 2002), they did not correlate rhythm and rate measures systematically (Grabe & Low, 2002; Dellwo & Wagner, 2003; Barry *et al.*, 2003; Ramus, 2002), or they only looked at very selected rhythmic measures (Grabe & Low, 2002; Dellwo & Wagner, 2003; Dellwo, 2006). The problem that rhythmic measures can interact with rate was addressed in other studies by suggesting rate normalization procedures for selected measures but they have not been seldom been without dispute. For example, Grabe & Low (2002) used a rate normalization method for vocalic rhythm measure (nPVI), but Barry *et al.* (2003) claims that the applied method does not fulfill its purpose. Dellwo (2006) introduced a rate normalization method for ΔC by measuring the standard deviation proportional to the mean (coefficient of variation). White & Mattys (2007) extended this technique to vocalic interval variability ΔV by measuring VarcoV.

[2] In case of %V the situation is slightly different. As a ratio measure it does not reveal the variability of V intervals. However, it is assumed to be influenced by variable consonantal complexity and vocalic reduction (high consonantal complexity may lead to a higher overall proportional content of C intervals, vocalic reductions may lead to shorter vowels and a lower overall V interval proportion; see Ramus *et al.*, 1999, for details).

Both Dellwo and White & Mattys found that this rate normalization can be of advantage, for example, as they found strong support for the rhythm class concept by the rate normalized measures. The present study, however, will reveal that the Varco-normalization can be problematic since the typical data distributions of C and V interval durations do not meet the assumptions underlying such calculations (see section 4). The same criticism applies for a rate normalization method for the PVI based on z-transformations of the raw data (Wagner & Dellwo, 2004).

Given the gap of systematic knowledge about the degree to which individual rhythm measurements are dependent on rate, and about the effectiveness of rate normalization procedures the present study has three aims:

(a) In experiment 1 (section 3) the influence of rate on the four most widely used rhythm measures %V, ΔC, consonantal PVI and vocalic PVI were studied.

(b) In experiment 2 (section 4) the effectiveness of rate normalization methods for rhythm measures was tested.

(c) In experiment 3 (section 5) it was tested how effectively rate normalized measures separate languages of different rhythm classes.

3. INFLUENCE OF RATE ON RHYTHM MEASURES

In the present experiment it was tested which of four widely used rhythm measures based on C and V interval variability (%V, ΔC, nPVI, rPVI) are influenced by rate variability of C and V data. To trigger effects of interval rate on the rhythm measures under investigation speech data was used in which speakers tried to vary their speech tempo from very slow to very fast, thus producing a wide range of interval rates.

3.1 Method

Speech Material: The speech material was taken from a database designed for speech rhythm and rate analysis at Bonn University and University College London (BonnTempo Corpus; Dellwo *et al.*, 2004; Dellwo, forthcoming). The speech material is based on a texts which is a short passage from a novel by Bernhard Schlink ('Selbs Betrug') with 76 syllables in the German version. This text was translated into the other languages under investigation by philologically educated native speakers of the target languages English (77 syllables), French (93 syllables), Italian (106 syllables). The languages were selected to represent both traditional rhythmic classes. 'Stress-timing' is represented by English and German, 'syllable-timing' by French and Italian and a language that is difficult to classify on an auditory basis, Czech.

The text was read by each speaker in each language under five different intended speech tempo versions. Speakers were first asked to read normal, then slow, then even slower (very slow), then fast, and then as fast as possible (see Dellwo *et al.*, 2004, for details of the recording procedure) leading to five different intended tempo categories from very slow to as fast as possible. This method resulted in a wide variety of C and V rates for each speaker and language recorded. Thus the data is highly suitable for studying effects of rate on rhythmic variability in speech.

language	speakers	syllables	C-intervals	V-intervals	pauses
English	7	2684	2475	2444	261
French	6	2734	2420	2455	250
German	15	5698	5028	4832	468
Italian	3	1619	1335	1380	95
Czech	8	3720	3608	3653	392
Total	**39**	**16455**	**14866**	**14764**	**1466**

Table 1: Number of languages, speakers, syllables, C intervals, V intervals, and pauses for native speakers of the languages English, French, German, Italian, and Czech in the dataset of the present study

Measures and measurement units: Four measures were tested for rate influences listed below. Each measure was calculated for an interval of speech between naturally occurring pauses (inter-pause-interval) so that no pause content was included in the calculations. Typically inter-pause-intervals consisted of a clause or a sub-clause but this may vary slightly between different speakers and it may vary tremendously between different intended speech tempi (at higher intended tempi fewer pauses were produced).

The measurements of rhythm and rate were:

(a) The percentage over which speech is vocalic (%V). This measure is a ratio measure showing the proportional differences between the overall durational vocalic and consonantal content in speech (see APPENDIX I, Equation 2 for a formula).

(b) The standard deviation of C interval durations (ΔC; APPENDIX I, Equation 3).

(c) The 'raw' consonantal Pairwise Variability Index (rPVI). This measure calculates the average difference between consecutive C interval durations in each inter-pause-interval (APPENDIX I, Equation 4).

(d) The normalized vocalic Pairwise Variability Index (nPVI). This measure calculates averages of relative differences between consecutive V intervals. Relative differences are obtained by calculating each difference proportional to the overall duration of the two consecutive V intervals under observation (APPENDIX I, Equation 5).

(e) Rate was measured as the average number of C and V intervals (combined) per inter-pause-interval.

Procedure: Four widely used measures of speech rhythm (see introduction) were first cross-plotted against the rate of C and V intervals (CV interval rate) for a descriptive analysis. With a curve analysis procedure it was then tested which mathematical function best describes the relationship.

3.2 Results

Figure 1 shows %V (top left), ΔC (top right), nPVI (bottom left), and rPVI (bottom right) cross-plotted against CV-rate. Descriptively the graphs reveal clearly that there is a relationship between CV-rate and ΔC as well as rPVI. In both cases the relationship can be described as a negative correlation, i.e. an increase in CV-rate leads to a decrease in ΔC or rPVI. No relationship can be detected between either %V or nPVI and CV-rate.

A linear and a logarithmic curve have been fitted to all four data plots. As can be expected from the descriptive analysis the R square for the %V/CV-rate and the nPVI/CV-rate comparisons turned out very poorly. For %V both the linear and logarithmic curve return a value of 0.035 ($p>0.05$). A very similar situation is the case for nPVI (R square

linear: 0.03, logarithmic: 0.031; p>0.05). It can thus safely be concluded that there is no influence of CV-rate on either %V or nPVI.

For ΔC the returned R square value for the linear fit results in 0.535 and for the logarithmic fit in 0.63 (p<0.005 in both cases). The situation is very similar for rPVI (R square linear: .492, logarithmic: .577; p<0.005). It can thus be concluded that both ΔC and rPVI are highly dependent on variability of CV-rate and that a logarithmic model is best for describing the relationship between the two parameters.

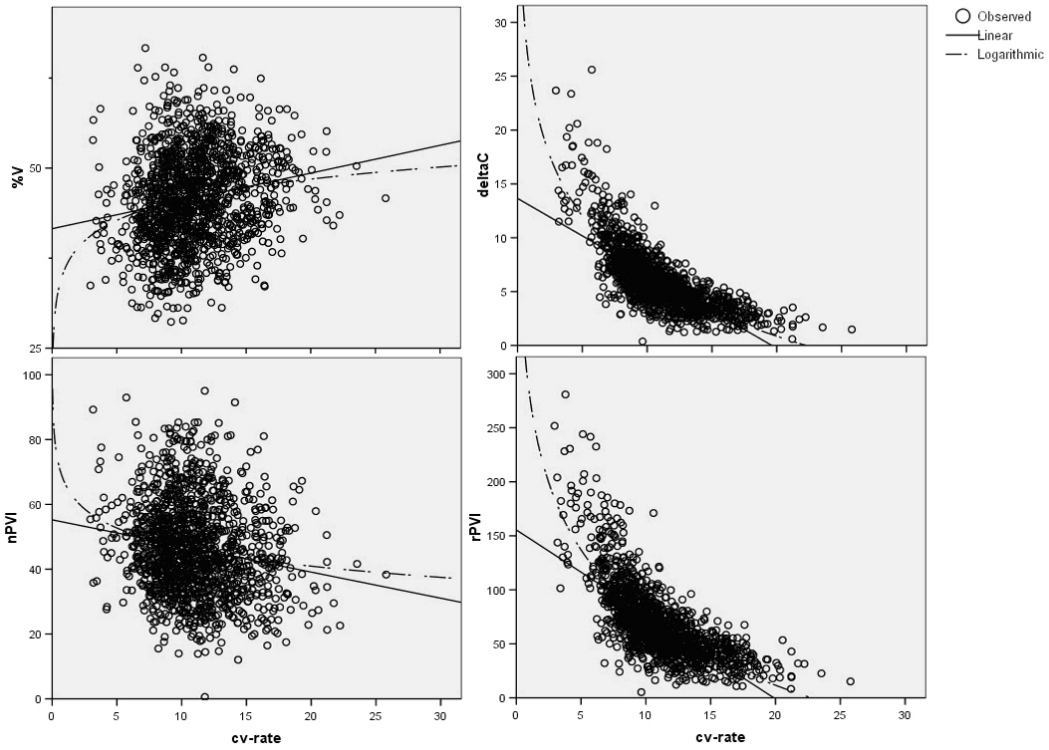


Figure 1: Scatter plot the rhythm measures %V (top left), ΔC (top right), rPVI (bottom left), and nPVI (bottom right) as a function of CV-rate with a linear and logarithmic curve fitted (each point is defined by the respective rhythm and rate values obtained from one inter-pause-interval)

3.3 Discussion

Both the V-interval variability measures %V and nPVI are clearly not dependent on CV-rate but the C interval measures rPVI and ΔC proved to be strongly affected. An explanation for this finding is straightforward: When speech rate is slower, C-intervals are longer and thus C-interval variability is higher. This affects the standard deviation of C-interval durations (ΔC) and the absolute durational variability monitored by the rPVI. The findings reveal that even in data with probably maximum CV-rate variability the measures %V and nPVI are not affected by the rate parameter. This demonstrates that the rate

normalization procedure applied for the nPVI (see Grabe & Low, 2002) is effective contrary to the concerns in Barry *et al.* (2003) where this normalization procedure was put into question (see Introduction).

The results also show clearly that a rate normalization is necessary both in case of ΔC and the rPVI. If these measures are not normalized they inevitably carry speech rate information thus it will be unclear to which extent obtained rhythmic differences between two groups may be the result of rate variability. Grabe & Low (2002) argue that rate normalization of the rPVI is problematic because rate differences in the rPVI will be a combined effect of speaking rate and between language differences in syllable structure. It remains unclear, however, why this interaction should prevent one from carrying out a normalization since both factors contribute to the same parameter: *rate*. After all it seems somehow irrelevant for measuring speech rhythm where rate influences derived from as long as they are in the signal. For this reason it is argued here that rPVI will need to be rate normalized if rhythmic measurements of c interval variability want to be obtained.

In the next section rate normalization methods will be applied for ΔC and rPVI and their effectiveness will be tested. After this (section 5) it will be tested how well the rate normalized ΔC and rPVI support the impression that languages of different rhythmic classes (stress- and syllable-timed) vary in rhythm.

4. NORMALIZING RHYTHM MEASURES FOR RATE

In this section appropriate rate normalization methods for ΔC and rPVI will be suggested and it will be tested how efficient they are by comparing the influence of speech rate on the measures before and after normalization.

4.1 Normalizing ΔC for rate

4.1.1 Data considerations for the calculation of ΔC

It was first tested whether the data assumptions are met for calculating ΔC. The calculation of a standard deviation (e.g. the standard deviation of C interval durations, ΔC) assumes the data to be normally distributed (Gaussian normal distribution). Thus, any study on speech rhythm based on the calculation of standard deviations of any interval in speech (C and V intervals, but also syllables or inter-stress intervals) needs to guarantee that the underlying data is normally distributed. However, interval durations in speech should not necessarily be assumed to be normally distributed. After all, speech segments have some threshold of maximum shortness (a segment can hardly be shorter than 5 ms) but there is probably no threshold limiting the length of a segment (especially vowels can be of very long duration, for example, as an effect of phrase final lengthening). For this reason it should be assumed that speech units of any type (consonants, C intervals, vowels, V intervals, syllables, etc.) may well be positively skewed, i.e. the right part of the data distribution graph possesses a long tail which is the result of a comparatively low frequency of data points of high durations. It should also be assumed that a higher proportion of values is distributed around the lower threshold of the data which leads to positive kurtosis. This assumption was tested by plotting the distributions for C, V and Syllabic (S) interval durations in the data set described in the previous section (BonnTempo Corpus; Dellwo *et al.*, 2004).

Results reveal that the assumption of non-normally distributed interval durations is supported by the data. Figure 2 (top) displays the distributions of C (left), V (right) and

Syllable (S; centre) durations. A black line superimposes a Gaussian normal distribution. It is clearly visible for each case that the bulk of the data is shifted to the left of the normal distribution peak and higher values occur at low frequency (positive Skew). Furthermore the peak values of the central data are much higher than for a normal distribution (positive Kurtosis).

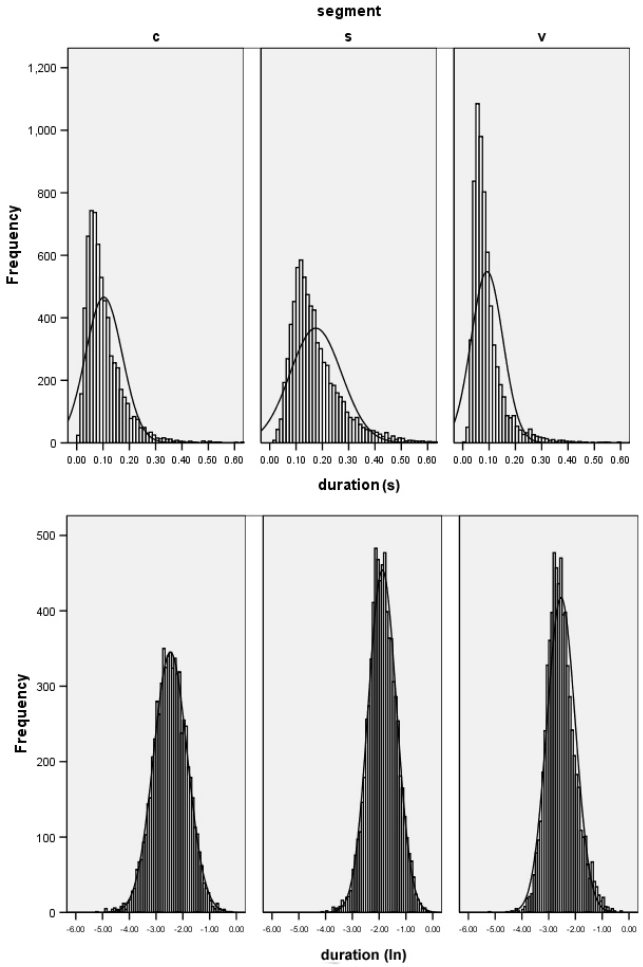


Figure 2: Histograms showing the distribution of C (left), V (right) and syllabic (centre) interval durations with superimposed Gaussian normal distribution for raw durations (top) and logarithmically transformed durations (bottom)

A common way of reducing positive Skew and Kurtosis is by calculating the logarithm for each interval duration (logarithmic transform, henceforth: ln transform), typically to the base e (Euler's number). Descriptive results of this transformation can be obtained from the histograms in Figure 2 (bottom). The figure contains the distributions for ln transformed C (left), S (centre) and V (right) intervals durations with superimposed

normal distribution lines. It is clearly visible that the interval duration distributions are now much closer to a Gaussian normal distribution and can thus be regarded as approximately normally distributed.

In addition to the descriptive analysis, Skewness and Kurtosis coefficients for the S, C and V intervals before and after the ln transformation were calculated and are obtainable from Table 2. The table displays the results for the raw (raw) and ln transformed data (ln). The table also contains the standard error for the deviation of ln transformed durations from the normal distribution. It is clear from the table that Skewness got significantly reduced from positive values between roughly 2 and 3 to values around 0. Only in the case of V-intervals a Skewness coefficient of .45 still remains. However, such an amount of Skewness is still in an acceptable range. The ln transform also had a tremendous effect on Kurtosis values which were reduced from values between about 5 and 14 to values not higher than 0.5. In the case of C intervals Kurtosis even came down close to 0 (0.084). The low standard errors (<0.034) for the ln data show that the ln transformed data distributions do not locally deviate much from the normal distributions.

	Skewness			Kurtosis		
unit	raw	ln	standard error	raw	ln	standard error
S	1.72	-.06	.016	4.8	.24	.032
V	2.87	.45	.017	14.38	.47	.033
C	2.11	-.01	.017	7.98	.084	.033

Table 2: Values for Skewness and Kurtosis before (raw) and after (ln) ln transformation for syllabic (S), consonantal (C) and vocalic (V) intervals (standard error refers to deviation of the ln transformed data distribution from a normal distribution)

It has been demonstrated in this section that syllabic as well as C and V intervals are strongly positively skewed and reveal a considerable amount of Kurtosis which would possibly strongly influence all analysis procedures assuming normally distributed data. This is clearly the case for ΔC as the calculation of a standard deviation assumes a normal (i.e. Gaussian) data distribution. It could be demonstrated that a logarithmic transformation of C-interval durations leads to a satisfactory normal distribution of the data.

4.1.2 Testing the influence of rate on ΔC based on ln transformed C durations

For this section ΔC was calculated based on ln transformed C interval durations (this measure is henceforth referred to as: ΔCln) and the influence of rate on this measure was tested with the same method as in section 3 (first, plotting ΔCln across CV-rate for a descriptive analysis and then correlating the two parameters to test the strength of the relationship). The results can be viewed in Figure 3 (left) which contains the cross-plot of ΔCln over CV-rate. The plot reveals that there is no obvious relationship between the two parameters. A linear regression analysis confirms this visual impression with a low and insignificant R-square value of 0.045 ($p>0.05$).

This result shows that the ln transform of the data is sufficient for normalizing CV-rate effects on ΔC. An explanation for the effect is straightforward. The ln transform leaves intervals of short durations at more or less equal duration whereas long durations are shortened drastically. In this respect, short durations from fast rates approximate long durations of slow rates. What remains is the proportional variability but no longer the high absolute interval differences.

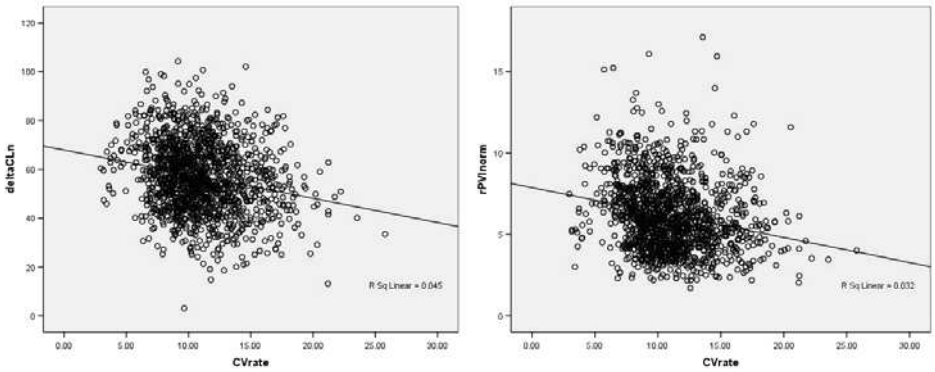


Figure 3: Scatter-plot showing ΔC based on ln transformed data (left) and rate normalized rPVInorm (right), both plotted over CV-rate

4.1.3 Discussion

In the previous two sections it was demonstrated, first, that the data distributions of C, V and S interval durations does not meet the assumptions necessary for the calculation of standard deviations (e.g. ΔC), second, that a logarithmic transform of interval durations changes the data distributions to fulfill the assumptions and, third, that ΔC based on ln transformed durations (ΔCln) is not speech rate dependent. As such, the ln transform of the data is a suitable rate normalization procedure for ΔC. Rate normalization methods based on the coefficient of variation of ΔC (varcoC =(ΔC*100)/meanC; Dellwo, 2006, White and Mattys, 2007) are based on the absolute standard deviation and it is questionable to what extent such measures are suitable for further statistical processing (e.g. for referential statistic methods assuming normally distributed data).

4.2 Normalizing rPVI for rate

4.2.1 Normalization procedure

Previous research applied rate normalization methods for rPVI by calculating the measures for z-transformed data (Wagner & Dellwo, 2004). However, also the z-transform assumes a normal distribution of the data which is not the case for C interval duration distributions (see above). A solution might be to apply the z-transform to logarithmically transformed durations. However, by performing a logarithmic transform of the data, large differences between consecutive C intervals become reduced drastically. It is probably counterproductive to reduce such differences as they are the differences that are the basis for rPVI variability. For this reason this technique was not further pursued in the present study.

In case of the PVI a rate normalization method already exists and is widely applied for the vocalic nPVI and it was demonstrated to be effective as the nPVI based on V interval durations revealed not to be dependent on rate (section 3). For this reason the same rate normalization method will be applied for rPVI. This is an easy procedure as only relative instead of absolute differences between consecutive C interval durations need to be averaged (see APPENDIX I, Equation 7). To avoid confusion with the vocalic nPVI measure the measure will be referred to as rPVInorm. This measure was correlated with CV-rate in the same way as ΔCln (above) using the same data as described in section 3.

4.2.2 Results

The results of the rate normalized consonantal rPVInorm can be viewed in Figure 3 (right plot). The graph shows the rPVInorm plotted across CV-rate and it is obvious that the normalization procedure is an effective control for speech rate in case of the consonantal variability measure. With an R square of 0.032 ($p>0.05$) as a result of a linear regression analysis it can be concluded that there is no correlation between the CV-rate and the rPVInorm.

4.2.3 Conclusions about rPVI normalization

It was demonstrated that the consonantal rPVI measure can be normalized effectively using the same rate normalization method as for the vocalic nPVI measure. Such a normalization method is probably more appropriate than existing methods using z-transform (Wagner & Dellwo, 2004) because the z-transform assumes normal distributions of the underlying data.

5. THE POWER OF RATE NORMALIZED RHYTHM MEASURES TO SEPARATE LANGUAGES OF DIFFERENT RHYTHMIC CLASSES

In section 3 it was demonstrated that the rhythmic correlates ΔC and nPVI correlate with the rate of C and V intervals. Section 4 showed how to normalize these measures for rate variability effectively. It remains to be tested whether the rate normalized measures still fulfill one of their main purposes, namely whether they support the auditory impression that languages of different rhythmic class sound different in their rhythm (more or less regularly timed; see introduction).

It is also possible that different languages react differently to rate normalizations. Languages with a low C interval complexity (e.g. French and Italian) may be able to maintain this complexity at high speeds while languages with a high C interval complexity (e.g. German and English) may reduce complex C intervals with increasing speed as a result of segment elision, for example. This may lead to a situation in which relative C interval variability measured by rate normalized measures (ΔCln and rPVInorm) changes with rate in some but not in other languages.

To compare speech rate influences on rhythm measures between languages, rates have to be made comparable across the different languages under investigation. For this reason rates in five parts of the total distributions (quintiles) were compared with each other. Quintiles were chosen because they roughly corresponded with the five intended tempo categories speakers produced. These intended tempo categories were not chosen as rate indicators as absolute speech rates within these categories may vary widely (see Dellwo, 2008)

5.1 Method

Speakers and Speech Data: The same data as in section 3 (Experiment I) was used in the present experiment.

Procedure: ΔCln and rPVInorm were calculated for each quintile of CV-rate for each of the five languages (Czech, English, French, German, Italian). Mean values of the results for each of the five languages were plotted across the five rate quintiles and ANOVAs were processed to analyze within and between language variability.

5.2 Results

Figure 4 contains the descriptive results for ΔCln (top) and rPVInorm (bottom) before (left) and after (right) normalization. Each graph contains the mean values for each respective measure and language at each quintile of CV-rate. Mean values were interpolated in each language with a line.

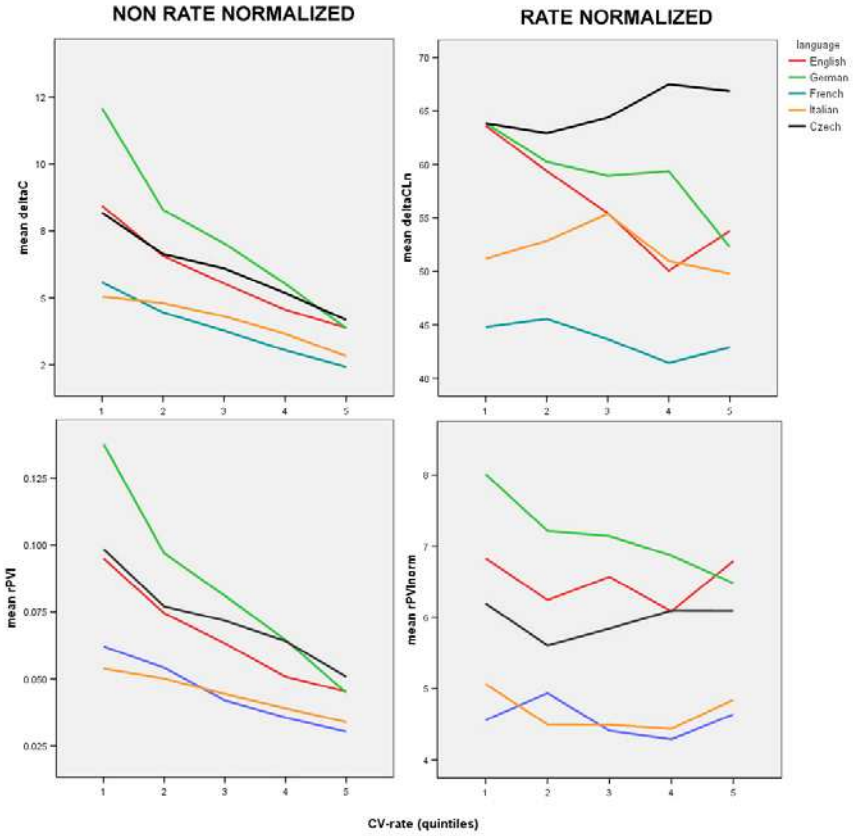


Figure 4: ΔC (top) and rPVI (bottom) before (left) and after (right) normalization. Graphs plot mean values for each quintile of CV-rate connected by a line for each of the five languages under investigation (Czech, English, French, German, Italian)

Both graphs on the left in Figure 4 reveal the effects of speech rated on the non normalized measures as it can be seen that both rhythm measures drop in each language with increasing CV-rate quintile. This should of course be expected given the analysis in section 3. However, upon visual inspection the results in both graphs look extremely similar and it is important to notice that at all CV-rate categories the rhythm class theory is well supported: the two syllable-timed languages, Italian and French, are both very similar and clearly lower than stress-timed English and German. Unclassified Czech is more similar to traditional stress-timed languages, English and German. This is true at all individual CV-rate quintiles. However, when looking at the results across the rate quintiles both graphs imply that the rhythm of slow French and Italian (1st quintile) is similar to medium speed English and Czech (3rd quintile) which again is similar to fast German (5th quintile). Such conclusions are likely to be incorrect because these similarities are a result of CV-rate variability in the data and not C interval durational variability.

After normalization (Figure 4, left graphs) the picture changes drastically and not many similarities between nPVInorm and ΔCln remain. In both cases it can be observed that speech rate normalization had an effect in that speech rate influences are either less systematic or not present any more. However, in particular in the case of ΔCln, strong differences can be observed between languages. Also, upon visual inspection the rhythm class distinction is still supported in the case of rPVInorm (French and Italian are both much lower than English and German) but not any more in the case of ΔCln where languages like Italian and English (syllable-timed and stress-timed) have about the same mean values for the 3rd and 4th quintile (which is probably the most common rate in each language). At the 5th quintile (fastest CV-rates) the languages German, English and Italian group together, suggesting that they share rhythmical features. Czech is drastically higher and French drastically lower than these languages, suggesting that they are rhythmically very different. At the 1st quintile Czech, English, and German are grouping up (again suggesting similarity in rhythm). Because of their different nature after normalization both measures, ΔCln and rPVInorm, will be discussed separately in the following.

rPVInorm: The effect of speech rate in the five different languages after the normalization has been tested with five ANOVAs (one for each language). Results show that there is only significant variability between the quintiles in the case of German ($F[4,524]=4.06$, $p= 0.003$) but not for any other language (English: $F[4, 524]=0.54$, $p=0.71$; French: $F[4, 209]=1.14$, $p=0.351$; Italian: $F[4, 104]=0.52$, $p=0.72$; Czech: $F[4, 314]=1.02$, $p=0.34$). A post-hoc analysis for German shows that significant variability of rPVI across the CV-rate quintiles is only existent between quintile 1 and 4 ($p=0.03$) and 1 and 5 ($p=0.001$). This analysis shows that speech rate normalization in the case of rPVInorm was very effective. Speech rate differences only remain in one langauge, German, and there only between rather extreme rates. These differences, however, can probably be neglected since the extreme rate ranges in BonnTempo are unlikely to occur regularly in real speech situations (Dellwo *et al.*, 2004).

Additional five ANOVAs tested rhythm class differences at each of the five quintiles with rPVI as dependent variable and rhythm class as a 2 class factor. Czech was excluded from this analysis as its rhythmic categorization has been disputed (Dancovicova & Dellwo, 2007); English and German were attributed to the stress-timed, French and Italian to the syllable-timed group. The analysis showed that at each quintile highly significant differences were obtainable between rhythm classes ($p<0.001$). It can therefore be

concluded that the rate normalized rPVInorm is a very robust variability measure supporting the rhythm class hypothesis across a range of extreme speech rates in all languages.

ΔCln: Rate normalization in the case of ΔC creates a very different picture. Languages seem to be very unequally affected by the rate normalization and a rhythm class distinction is not very well obtainable any more. ANOVAs for testing within-language variation (quintile as a five class factor and language as the dependent variable) help interpreting the situation and show that the visual impression may be a slightly misleading: Only for the languages English and German a significant variability between the quintiles can be detected (English: $F[4,244]=6.5$, $p<.001$; German: $F[4, 524]=12.1$, $p<.001$) but not for any of the other languages under investigation (French: $F[4,209]=.99$, $p=.42$, Italian: $F[4, 104]=.82$, $p=.51$, Czech: $F[4, 314]=1.7$, $p=.143$). Post-hoc the analysis reveals for English that there is significant difference between the quintile pairs 1/3, 1/4, and 1/5 as well as pair 2/4. In German only the 5th quintile is significantly different from all other groups. Given these results, it can be concluded that rate normalization is probably not as ineffective as the visual impression of the data suggests (apart from a few exceptions in English and German). However, the support for rhythm classes in this measure is less strong. More research needs to be performed to ΔCln to produce a clearer picture.

5.3 Discussion and conclusion

It can be concluded that the rate normalization procedures applied to rPVI and ΔC lead to a more robust picture of rate. However, because of these traces rhythmic class separation of ΔC is somehow distorted after normalization. For this reason rPVI is considered the more robust measure for rhythmic class separation when rate variability is present in the data.

Given the assumption above (5) that the complexity of C intervals in stress-timed languages may favor variability of C interval durations across rates (because complex intervals are likely to be reduced in complexity at higher rates) we found support here only with ΔCln (ANOVAS for within language variability across rates revealed significant differences for English and German) but not with rPVInorm. So the raised concerns in Grabe & Low (2002) against normalizing rPVI because between language syllable complexity differences might interact with speakers' speech rate (see Introduction) seem unjustified in the light of the present results.

6. OVERALL SUMMARY AND CONCLUSIONS

In this paper it was demonstrated that the widely used rhythm measures ΔC and rPVI correlate with CV-rate variability and that %V and nPVI are unaffected by rate. It was demonstrated that there are effective ways to normalize the rate affected measures ΔC and rPVI. For the rhythm measures' power to separate languages of different rhythm classes on an acoustic level it was demonstrated that the normalized rPVI shows more consistent results than the normalized ΔC.

The main conclusions of this paper are thus that both consonantal rhythm measures, ΔC and rPVI, need to be normalized for rate when C interval variability is intended to be measured. If the aim of the rhythm analysis is to separate languages of different rhythm classes from each other then rPVInorm is probably the best choice when a rate normalized consonantal measure needs to be chosen.

7. REFERENCES

Abercrombie, D. (1967), *Elements of General Phonetics*, Edinburgh: University Press.

Bolinger, D.L. (1981), *Two kinds of vowels, two kinds of rhythm*, Bloomington, Indiana: Indiana University Linguistics Club.

Barry, W. J., Andreeva, B., Russo, M., Dimitrova, S., and Kostadinova, T. (2003), Do rhythm measures tell us anything about language type?, in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2693–2696.

Cahn, J. E. (1990), The generation of affect in synthesized speech, in *Journal of the American Voice I/O Society*,
(electronic Publication: https://eprints.kfupm.edu.sa/70011/1/70011.pdf)

Classe, A. (1939), *The rhythm of English prose*, Oxford: Blackwell.

Cutler, A. (1997), The syllable's role in the segmentation of stress languages, in *Language and Cognitive Processes*, 12, 839-845.

Cutler, A. & Norris, D. G. (1988), The role of strong syllables in segmentation for lexical access, in *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113-121.

Dancovicova, J. & Dellwo, V. (2007), Czech speech rhythm and the rhythm class hypothesis, in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, 1241-1244.

Dauer, R.M. (1983), Stress-timing and syllable-timing reanalyzed, *Journal of Phonetics*, 11, 51-69.

Dauer, R.M. (1987), Phonetic and phonological components of language rhythm, in *Proceedings of the 11th International Congress of Phonetic Sciences*, Talinn, Estonia, 447-450.

Dellwo, V. (forthcoming), *Influences of speech rate on acoustic correlates of speech rhythm: An experimental investigation based on acoustic and perceptual evidence*, PhD thesis, Bonn University, Germany.

Dellwo, V. (2006), Rhythm and Speech Rate: A Variation Coefficient for ΔC, in *Language and Language-processing* (P. Karnowski & I. Szigeti, editors), Frankfurt am Main: Peter Lang, 231-241.

Dellwo, V. (2008), The role of speech rate in perceiving speech rhythm, in *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 375-378.

Dellwo, V. & Wagner, P. (2003), Relations between Language Rhythm and Speech Rate, in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 471–474.

Dellwo, V. & Koreman, J. (2008), How speaker idiosyncratic is acoustically measurable speech rhythm?, in *Electronic Proceedings of the annual meeting of the International Association of Forensic Phonetics and Acoustics (IAFPA)*, Lausanne, Switzerland.

Deterding, D. (2001), The measurement of rhythm: A comparison of Singapore and British English, *Journal of Phonetics,* 29, 217-230.

Grabe, E. and Low, E. L. (2002), Durational variability in speech and the rhythm class hypothesis, in *Papers in Laboratory Phonology 7* (C. Gussenhoven & N. Warner, editors), Berlin: Mouton de Gruyter.

James, A. L. (1929), *Historical introduction to French Phonetics*, London: ULP.

Kim, J., Davis, C., and Cutler, A. (2008), Perceptual tests of rhythmic similarity: II. Syllable rhythm, *Language and speech*, 51, 343-359.

Low, E.L., Grabe, E. & Nolan, F. (2000), Quantitative characterization of speech rhythm: Syllable-timing in Singapore English, *Language and Speech*, 43, 377-401.

Mok, P. & Dellwo, V. (2008), Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English, in *Proceedings of Speech Prosody*, Campinas, Brazil, 423-426.

Nazzi, T., Bertocini, J. & Mehler, J. (1998), Language discrimination by newborns: Towards an understanding of the role of rhythm, *Journal of Experimental Psychology: Human Perception and Performance*, 24, 756-766.

Pike, K.L. (1945), *Intonation of American English*, Ann Arbor: University of Michigan Press.

Ramus, F. (2002), Language discrimination by newborns, *Annual Review of Language Acquisition*, 2, 85-115.

Ramus, F., Hauser, M.D., Miller, C., Morris, D. & Mehler, J. (2000), Language discrimination by human newborns and cotton-top tamarin monkeys, *Science*, 288, 349-351.

Ramus, F. & Mehler, J. (1999), Language identification based on suprasegmental cues: A study based on resynthesis, *Journal of the Acoustical Society of America*, 105(1), 512-521.

Ramus, F., Nespor, M. & Mehler, J. (1999), Correlates of linguistic rhythm in the speech signal, *Cognition*, 73, 265-292.

Roach, P. (1982), On the distinction between 'stress-timed' and 'syllable-timed' languages, in *Linguistic controversies* (D. Crystal, editor), London: Edward Arnold, 73-79.

Rincoff, R., Hauser, M., Tsao, F., Spaepen, G., Ramus, F. & Mehler, J. (2005), The role of speech rhythm in languages discrimination: further tests with a non-human primate, *Developmental Science*, 8(1), 26-35.

Tilsen, S. & Johnson, K. (2008), Low-Frequency Fourier analysis of speech rhythm, *Journal of the Acoustical Society of America*, 124(2), (Online EL Publication).

Toro, J.M., Trobalon, J.B. & Sebastian-Galles, N. (2003), The use of prosodic cues in language discrimination tasks by rats, *Animal Cognition*, 6(2), 131-136.

Wagner, P. & Dellwo, V. (2004), Introducing YARD (Yet Another Rhythm Determination) and Re-Introducing Isochrony to Rhythm Research, in *Proceedings of Speech Prosody 2004*, Nara, Japan.

Wang, Q. (2008), L2 stress-perception: The reliance on different acoustic cues, in *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 635-638.

White, L. & Mattys, S. (2007), Calibrating rhythm: first language and second language studies, *Journal of Phonetics*, 35, 501-522.

White, L., Mattys, S., Series, L. & Gage, S. (2007), Rhythm metrics predict rhythmic discrimination, in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, 1009-1012.

APPENDIX I: LIST OF FORMULAS FOR THE MEASUREMENTS OF SPEECH RATE AND RHYTHM

Equation 1: *Combined C and V interval rate*

$$CVrate = \frac{n_C + n_V}{\sum_{i=1}^{n_C} c_i + \sum_{i=1}^{n_V} v_i}$$

n = number of sampled intervals
C = C interval
V = V interval
c = C interval duration
v = V interval duration

Equation 2: *Percentage over which speech is vocalic (%V)*

$$\%\mathrm{V} = \frac{\left(\sum_{i=1}^{n_v} v_i\right) \cdot 100}{\sum_{i=1}^{n_c} c_i + \sum_{i=1}^{n_v} v_i}$$

n_v = total number of V-interval samples
n_c = number of C-interval samples
v = V interval duration
c = C interval duration

Equation 3: *Standard deviation of C intervals (ΔC)*

$$\Delta C = 100 \cdot \sqrt{\frac{n \cdot \sum_{i=1}^{n} C_i^{\,2} - \left(\sum_{i=1}^{n} C_i \right)^2}{n \cdot (n-1)}}$$

n = number of sampled intervals
C = duration of C interval

Equation 4: *Non-normalized consonantal Pairwise Variability Index*

$$\mathrm{rPVI} = \frac{\sum_{c=1}^{n-1} \left| x_c - x_{c+1} \right|}{n-1}$$

n = number of C-intervals sampled
c = C interval duration

Equation 5: *Normalized vocalic Pairwise variability index*

$$\mathrm{nPVI} = 100 \cdot \frac{\sum_{v=1}^{n-1} \left| \frac{x_v - x_{v+1}}{(x_v + x_{v+1})/2} \right|}{n-1}$$

n = number of V-intervals sampled
v = V interval duration

Equation 6: *Coefficient of variation (varcoC) of ΔC*

$$\mathrm{varcoC} = \frac{\Delta c \cdot 100}{mean_C}$$

c = C interval duration

Equation 7: *normalized rPVI*

$$\mathrm{rPVInorm} = 100 \cdot \frac{\sum_{C=1}^{n-1} \left| \frac{x_c - x_{c+1}}{(x_c + x_{c+1})/2} \right|}{n-1}$$

n = number of C-intervals sampled
c = C interval duration

SPEECH RHYTHM AND WORD SEGMENTATION: A PROMINENCE-BASED ACCOUNT OF SOME CROSSLINGUISTIC DIFFERENCES

Christopher S. Lee
Goldsmiths, University of London
chrisslee@ntlworld.com

1. SUMMARY

Most work on crosslinguistic rhythmic differences has focused exclusively on temporal factors: the temporal distribution of phonological units at various levels in the classical timing-based accounts, or more recently the durational variability of phonetic units. Lee & Todd (2004) present a prominence-based account of speech rhythm, according to which a crucial determinant of rhythmic organisation is the variability in the auditory prominence of phonetic events (in particular vowels), as primarily determined by their duration, intensity and F0. According to this view, the key difference between so-called 'stress-timed' languages (e.g. English and Dutch) and other types is the greater variability in the prominence of their syllabic nuclei. They describe an auditory model developed by Todd and his associates and propose several prominence measures, all of which yield results consistent with their claim on the two multi-language corpora investigated.

Can such an account offer a possible explanation of why infant learners of languages such as English, as opposed to those of languages such as French, adopt a word-segmentation strategy based on the principle that a stressed syllable marks the likely boundary of a content word (Nazzi *et al.,* 2006)? The following claim is advanced here: perceptible prominence distinctions between neighbouring syllables (marking probable transitions between stressed and unstressed syllables) are sufficiently frequent in languages such as English (and Dutch) to delimit large numbers of learnable proto-words (i.e. no more than 2-3 syllables in length), whereas in languages such as French (and Italian), they are too sparsely distributed to serve a similar useful function. The results of analyses of the two multi-language corpora investigated in Lee & Todd (2004) are presented in support of the claim: they yield large and robust differences between the putative rhythm classes.

2. INTRODUCTION

There has been renewed interest in the last 10 years or so in the question of crosslinguistic differences in rhythmic structure and how such differences arise at the acoustic/auditory level. Classical timing-based accounts, going back to Abercrombie (1967), Ladefoged (1975), and Pike (1945), have given way to accounts based on differences in syllable structure, most notably by Ramus and his colleagues (Ramus & Mehler, 1999; Ramus *et al.*, 1999), and Grabe and her colleagues (Grabe & Low, 2002; Grabe *et al.*, 1999; Low *et al.*, 2000). Despite differences of detail, their work has converged on the conclusion that languages differ crucially in the degree to which their consonantal and vocalic intervals (uninterrupted stretches of uniquely consonantal or vocalic material) vary in duration, and also in the extent to which vocalic or consonantal material is predominant. They and other researchers have now adduced a large body of broadly supportive empirical evidence, ranging from instrumental studies showing how

their proposed metrics successfully categorise languages according to their generally agreed traditional rhythm class to perceptual studies showing how infant and adult listeners are able to distinguish resynthesised sentences (stripped of segmental and intonational cues) from languages that yield very different values on all/some of the metrics but are unable to distinguish those from languages that do not (see White & Mattys, 2007, for a review).

None of this work, however, sheds much light on the question of how such differences in rhythmic structure affect the way in which listeners process particular languages. A large body of research over the last 20 years or so has demonstrated the importance of rhythm for speech segmentation, and has highlighted crosslinguistic differences in the segmentation cues used by listeners. In particular, work by researchers such as Cutler and her colleagues on word-segmentation has shown that adult listeners make language-specific assumptions about the likely location of word-boundaries, with native listeners of languages such as Dutch and English, for example, but not French or Japanese, assuming that stressed syllables mark (content) word-onsets (Cutler *et al.*, 1986; Cutler & Norris, 1988; Cutler & Butterfield, 1992; Otake *et al.*, 1993; Vroomen *et al.*, 1996). More recently, the emergence of such differences in prelinguistic infants has been documented in studies of English, Dutch, French, and German learning infants: Dutch, English and German infants segment bisyllabic words beginning with a stressed syllable by the age of 8 months (Höhle *et al.*, 2001; Houston *et al.*, 2000; Juszcyk *et al.*, 1999), but (Parisian) French infants fail to segment bisyllabic words until after the age of 12 months, when they are able to make use of alternative non-rhythmic segmentation cues (Nazzi *et al.*, 2006).

What triggers the adoption of a stress-based segmentation strategy in infant learners of languages like English but not languages like French? In what follows, I sketch a possible answer. It is based on an account of speech rhythm proposed in Lee & Todd (2004), according to which a crucial difference between languages lies in the variability in prominence of their syllabic nuclei: their claim is that languages traditionally described as stress-timed, such as English, display greater variability in the prominence of their syllabic nuclei than languages traditionally described as syllable-timed, such as French, or mora-timed, such as Japanese. I first outline their account, and the supporting experimental evidence, before describing the proposal and the results of further analyses of the two multi-language corpora investigated in the original study.

3. A PROMINENCE-BASED ACCOUNT OF SPEECH RHYTHM

3.1 Introduction

Communicative auditory signals, such as music or speech, have a highly organised rhythmic structure. Their basic elements (typically notes/chords in music, syllables in speech) are grouped, on the basis primarily of their temporal distribution and perceptual prominence, into hierarchical layers of units, accompanied in the case of periodic (usually musical) input by a metrical framework (see Clarke, 1999; Lerdahl & Jackendoff, 1983; Patel, 2008). Prominent events are heard as *accented* in the context of a musical signal (Lerdahl & Jackendoff, 1983), or probably *stressed* in the context of a speech signal (see Ladd, 2008, on the relationship between stress and phonetic/acoustic cues), or simply *loud* where their prominence is owed to straightforwardly physical factors (usually intensity); what characterises all prominent events, however, is the fact that they capture the listener's attention (see Jones, 1987). The determinants of prominence are multiple and varied (as attested by work on musical rhythm perception; see e.g. the review in Parncutt & Drake,

2001), but the primary low-level determinants are intensity, duration and F0 (Plack & Carlyon, 1995).

Although past work on crosslinguistic rhythmic differences has focused almost exclusively on temporal factors, the notion that variability in prominence might be a significant crosslinguistic factor is not new (see e.g. Delattre, 1966, on the greater variability in intensity of English and German compared to French and Spanish vowels). The account proposed in Lee & Todd (2004) builds on such findings, and makes crucial use of the auditory model proposed by Todd and his colleagues (Todd, 1994; Todd & Brown, 1996; Todd *et al.*, 1999; Todd *et al.*, 2002). In the next two sections, I give an overview of the model, their account of crosslinguistic rhythmic differences, and the experimental evidence.

3.2 The 'rhythmogram' model

The model takes an audio signal as input and outputs an inferred sequence of primitive events together with their associated prominence values (it can also infer grouping boundaries and perform beat-induction, but I do not consider these aspects of the model here). It consists of 3 main components:

- An auditory periphery, consisting of an outer/middle ear filter, gammatone filterbank (modelling the frequency-selective properties of the basilar membrane), and an inner hair-cell model.
- A multi-scale low-pass mechanism (a bamk of low-pass filters, 12 per octave in the range 0.5-20Hz), which detects changes in the amplitude of a signal over a range of time scales.
- A peak detection and summation mechanism, which detects peaks in the outputs of the low-pass filters and sums peaks associated with a particular event to yield the event's prominence value *P*.

The output of the model may be conveniently visualised in the form of a graph of *P*-values and accompanying scatter-plot of peaks (with filter time-constant on the *y* axis and time on the *x* axis), the so-called 'rhythmogram' (see Figure 1).

It is beyond the scope of the current paper to give a detailed account of the model (for a fuller account, see Lee & Todd, 2004; Todd, 1994; Todd & Brown, 1996). But it is worth noting here how the model captures the primary low-level determinants of prominence. A more intense input event will cause the hair-cell model to "fire" more rapidly, thus resulting in a higher total spike rate (the units of *P*). Changes in F0 will reduce the effect of hair-cell adaptation, and hence result in an increased response to the event with the different F0, while frequency components in a sound event below 1-2kHz will be progressively attenuated by the outer/middle ear filter. Finally, events of longer duration will be associated with a longer uninterrupted low-pass response and hence higher *P*-values (the accumulation of peak responses, which is spread over time, continues until a new event occurs; see Lee & Todd, 2004).

Figure 1 shows the response of the model to one of the English sentences used in the experiments to test the theory (see next section). As can be seen, 9 of the 13 events contain the 9 syllabic nuclei (the other 4 events contain only consonantal segments). Most importantly, the analysis is sensitive to the distinction between stressed and unstressed syllables: the 4 most prominent events are those containing the 4 stressed syllable nuclei (events 2, 5, 8, and 12) while the unstressed syllables all yield weaker events (events 1, 4, 7, 10, and 11).

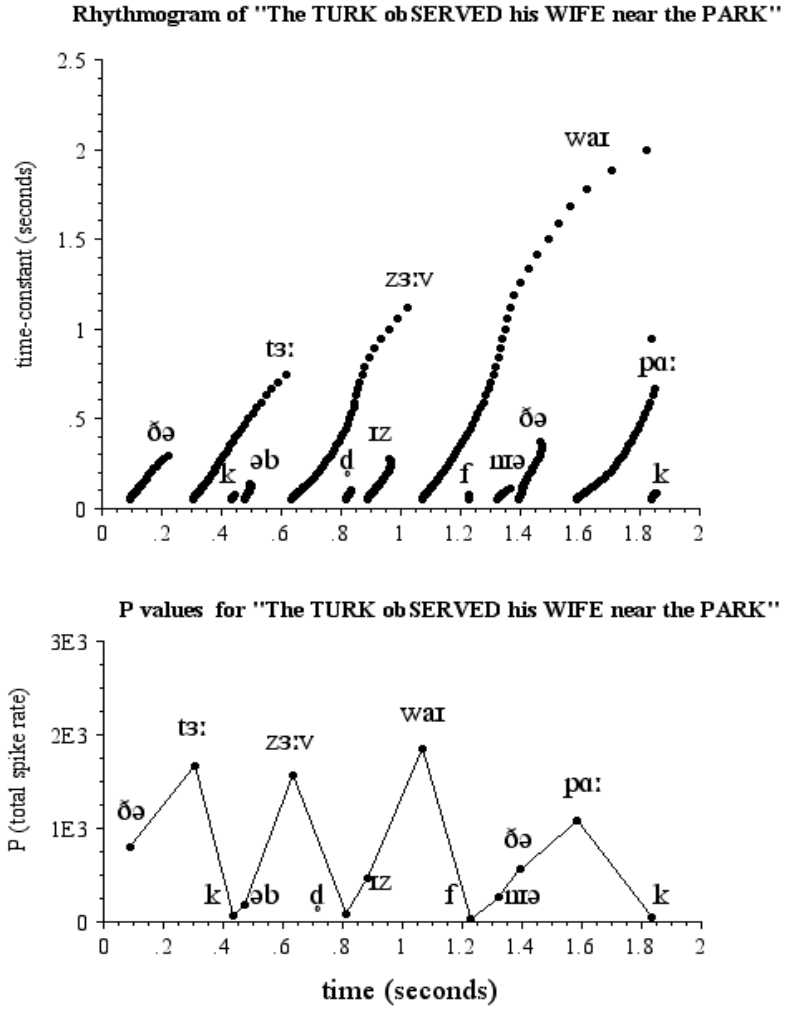


Figure 1: Rhythmogram representation (top) and corresponding P-values (bottom) for the sentence 'The Turk observed his wife near the park', read by a female speaker. From Lee & Todd (2004)

3.3 Crosslinguistic differences:hypotheses and experimental evidence

The claim advanced in Lee & Todd (2004), as already noted, is that languages like English display a greater variability in the prominence of their syllabic nuclei than languages like French (or Japanese). In order to test the claim, they ran the model on two multi-language corpora:

•A corpus of 5 English sentences (recorded by 4 male and 4 female native speakers) and 5 French sentences (recorded by 4 male and 4 female native speakers), matched for

number of syllables, number and location of stressed syllables, syntax and general meaning (the Lee & Todd - LT - sentences; see Appendix).

•A corpus of Dutch, English, French, and Italian sentences (recorded by 4 female native speakers of each language), drawn from a larger multi-language corpus recorded by Nazzi *et al.* (1998) and used by Ramus *et al.* (1999) in their study (the RNM sentences; see Nazzi *et al.*, 1998).

The main model-based measures were ΔP_{syll}, the standard deviation of the syllabic event prominences (events containing all/the most prominent part of a syllabic nucleus), and rPVI(P_{syll}), the raw pairwise variability index of syllabic event prominences. For comparison, they also derived F0 variability measures (pitch range and syllabic nuclear F0 variability) and 2 intensity measures (note that all sentences were normalised to a standard mean rms value before analysis; see Lee & Todd, 2004): ΔI, the standard deviation of the peak intensity values of syllabic nuclei, and ΔI_{adj}, the standard deviation of the intensity values adjusted for duration (I was decremented by 6dB for every halving of syllabic nuclear duration under 300ms in order to roughly model the effect of temporal integration). Finally they also computed all the main proposed measures of consonantal and vocalic variability (ΔC, ΔV, rPVI(C), nPVI(V), as well as varco(V) (the coefficient of variation of vocalic intervals) and %V.

The predictions were that the prominence variability measures (model-based, intensity and F0) would yield significantly higher values on the English (and Dutch) sentences than the French (and Italian) sentences. The results were as follows:

•For the LT sentences, all the measures (except %V and the F0 variability measures) yielded a statistically significant separation of the two languages in the predicted direction, with nPVI(V), ΔI_{adj}, and rPVI(P_{syll}) yielding the largest effects.

•For the RNM sentences, all the measures (except the F0 variability measures) yielded a statistically significant separation of the two language classes in the predicted direction, with nPVI(V), ΔI, and ΔI_{adj} yielding the largest effects.

In sum, both the model-based measures and the intensity measures performed as well as or better than the phonetic measures in separating the two sets of languages, suggesting that variability in prominence is indeed a significant crosslinguistic rhythmic factor. In the next section, I turn to the question of how such differences might trigger different lexical segmentation strategies in infant learners.

4. SPEECH RHYTHM AND LEXICAL SEGMENTATION

4.1 Hypothesis

An infant's adoption of a stress-based word segmentation strategy follows from his/her inference that there is a correlation between the acoustic cues signalling stress and the position of stressed syllables in words (Thiessen & Saffran, 2007). Like other probabilistic strategies adopted by infants in their search for words, such as those involving the calculation of transition probabilities between segments (eventually yielding the native-language phonotactics; see e.g. Jusczyk *et al.*, 1994) and syllables (see e.g. Saffran *et al.*, 1996), it depends crucially on the observation of frequencies of occurrence of phonetic/acoustic features in the linguistic input, and the assignment of frequently occurring and rarely occurring features to different cue-status categories. Hence, other things being equal, if there are only sparsely distributed stress cues in the linguistic input, infants are unlikely to adopt stress as a word-boundary cue; the contrary is true if stress

cues are relatively common. There are no findings bearing directly on the question of what might count as common enough, but given that very young (English-learning) infants seem to maximally segment trisyllabic words from continuous speech (Houston *et al.*, 2004) and that shorter words form the overwhelming preponderance of their vocabularies, it seems reasonable to assume that the occurrence of stressed syllables on average every 2 or 3 syllables would trigger the adoption of such a strategy, but their occurrence on average only half as frequently would probably fail to do so.

The hypothesis advanced here is that stress cues (in the form of prominence distinctions between neighbouring syllables) are sufficiently frequent on average in English and Dutch utterances to trigger a stress-based strategy, whereas in French and Italian utterances, despite the predictability of word-stress (at least in French), they are too sparsely distributed to do so. In the following section, I describe an experiment to test the hypothesis.

4.2 Experiment

4.2.1 Method

The two corpora used in the experiment were the LT sentences (see Appendix) and the RNM sentences (see Nazzi *et al.*, 1998). Prominence differences between neighbouring syllables in all sentences (the absolute differences between their values) were calculated on the model-based measure P_{syll}, and the two intensity measures, I and I_{adj} (in this case I was decremented by 3dB for every halving of duration under 150ms as a possibly more accurate reflection of psychophysical findings; see Plack & Carlyon, 1995).

There are clearly lower limits to the perceptibility of differences along the various acoustic dimensions that cue prominence. What are the JNDs (for adults and more particularly for infants) along the dimensions measured by the proposed metrics? For I, the question has a fairly straightforward answer: for adults the JND for intensity is 1-3dB, depending on the task and stimuli, whereas for infants it is considerably higher at around 6-9dB (see the review in Saffran *et al.*, 2006). For the other two measures, it is far from clear what the JNDs might be, so I make the simplifying assumption that they are equivalent to the JND for intensity. Hence for all three measures, the JND is set at 6dB (for P, this corresponds to a difference in total spike rate of 600, as determined by intensity scaling measurements on the model; see Lee & Todd, 2004).

It is clear that prominence differences below threshold are irrelevant to the perception of stress. Hence the dependent variable is the proportion of prominence differences that are at least 1 JND in magnitude.

4.2.2 Results and discussion

The results are shown in Figure 2. The hypothesis is supported in both corpora: the English and Dutch sentences yield a considerably higher proportion of prominence differences greater than 1 JND (on all three measures) than the French and Italian sentences (all language/language-class differences significant at $p < 0.001$with χ^2 on 1df). The results with I_{adj} and P_{syll} suggest that such prominence differences occur on average every 3 syllables or so in English and Dutch, compared to only every 4-5 syllables in French and Italian.

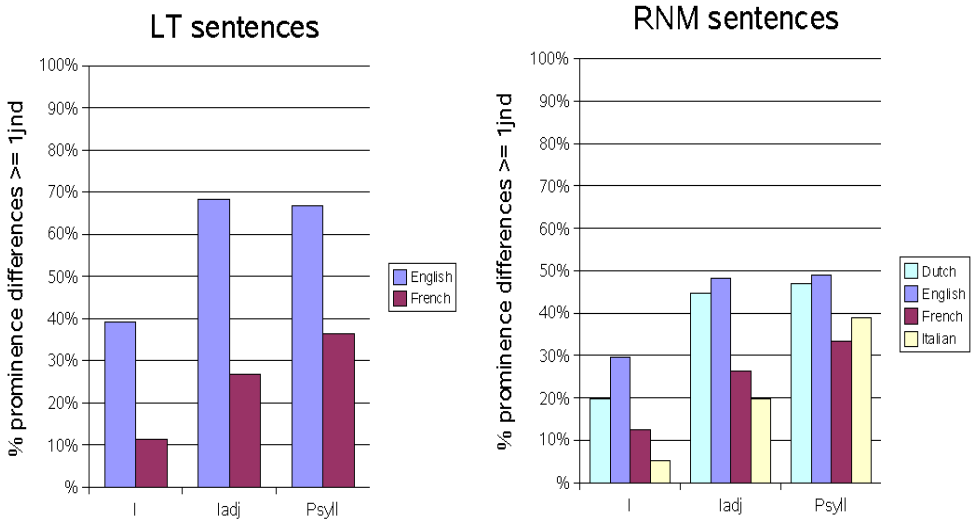


Figure 2: The percentage of prominence differences greater than 1 JND in the LT sentences (left) and the RNM sentences (right)

The language differences are particularly striking in the case of the LT sentences, since the sentences are exactly matched for the number and distribution of stressed and unstressed syllables. The results suggest that stress in French is less reliably cued by intensity/duration/F0 than in English. We can test this supposition further by looking at the closeness of the relationship between stress and prominence in the two sets of sentences. What proportion of stress differences (pairs of consecutive syllables that differ in stress) are marked by appropriate prominence differences (stressed syllable nucleus at least 1 JND more prominent than unstressed syllable nucleus) and what proportion of prominence differences of at least 1 JND appropriately mark stress differences?

The results are shown in Figure 3. As expected, the proportion of stress differences appropriately marked by prominence differences is very much higher in the English sentences than the French sentences (all language differences significant at $p < 0.001$with χ^2 on 1df). Furthermore, although prominence differences in both sets of sentences usually mark stress differences, they do so slightly more frequently in the English sentences than the French sentences (all language differences significant at $p < 0.05$ with χ^2 on 1df).

In sum, the results confirm that prominence cues likely to be large enough to be perceptible to infants are more sparsely distributed in French and Italian than English and Dutch, and that in French at least, one important reason is simply that stress differences are less reliably cued by prominence differences than in English.

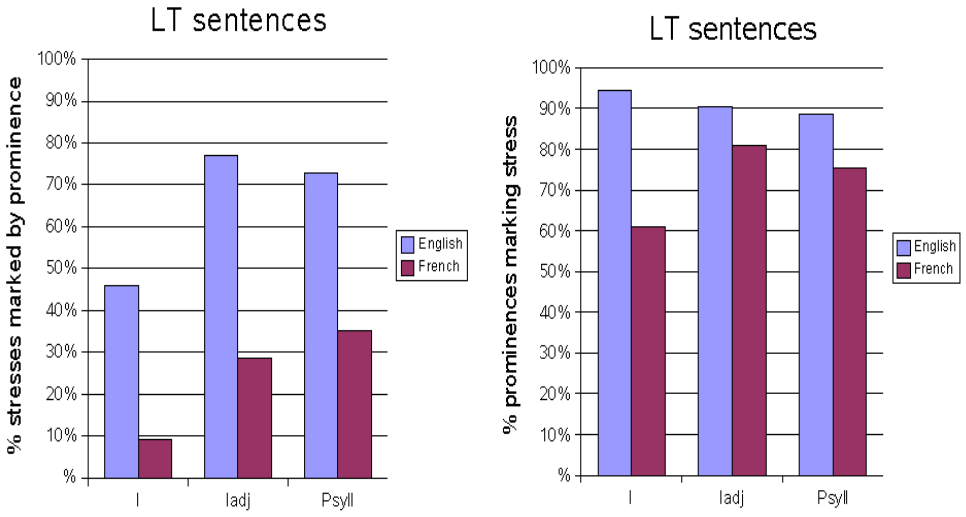


Figure 3: The percentage of stress differences appropriately marked by prominence differences (left) and the percentage of prominence differences appropriately marking stress differences (right) in the LT sentences

5. CONCLUSIONS

The prominence-based account of speech rhythm proposed in Lee & Todd (2004) captures an important crosslinguistic difference in rhythmic structure, as their experimental results demonstrate. In addition, as shown here, it yields a possible account of one way in which the rhythmic structure of a language seems to influence the process of lexical segmentation: infant learners of languages like English and Dutch, but not French and Italian, adopt stress as a lexical segmentation cue because prominence cues signalling probable stress contrasts are much more frequent in their linguistic input. It will be interesting to see if the account can be substantiated by further instrumental studies, which will need to include a range of different psychophysical assumptions (regarding e.g. JNDs) as well as further data-sets (and more languages) and perhaps also samples of infant-directed speech.

ACKNOWLEDGMENTS

Thanks to Neil Todd for his comments on the paper, and to the organisers of the AISV conference for giving me the opportunity to present this work at the roundtable on "Different ways of analyzing speech rhythm". An earlier version of this paper was presented as a poster at the Workshop on Empirical Approaches to Speech Rhythm, University College London, London, U.K., on 28 March 2008.

6. REFERENCES

Abercrombie, D. (1967), *Elements of general phonetics*, Edinburgh: Edinburgh University Press.

Clarke, E. (1999), Rhythm and timing in music, in *The psychology of music,* 2nd edition (D. Deutsch, editor), London: Academic Press, 473-500.

Cutler, A. & Butterfield, S. (1992), Rhythmic cues to speech segmentation: evidence from juncture misperception, *Journal of Memory and Language*, 31, 218-236.

Cutler, A., Mehler, J., Norris, D. & Segui, J. (1986), The syllable's differing role in the segmentation of French and English, *Journal of Memory and Language*, 25, 385-400.

Cutler, A. & Norris, D. (1988), The role of strong syllables in segmentation for lexical access, *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113-121.

Delattre, P. (1966), A comparison of syllable length conditioning among languages, *International Review of Applied Linguistics*, 4(3), 183-198.

Grabe, E. & Low, E.L. (2002), Durational variability in speech and the rhythm class hypothesis, in *Papers in laboratory phonology 7* (C. Gussenhoven & N. Warner, editors), Berlin: Mouton de Gruyter, 515-546.

Grabe, E., Post, B. & Watson, I. (1999), The acquisition of rhythm in French and English, in *Proceedings of the 14th International Congress of Phonetic Sciences*, 2,1201-1204.

Höhle, B., Giesecke, D. & Juszcyk, P. (2001), Word segmentation in a foreign language: further evidence for crosslinguistic strategies, *Journal of the Acoustical Society of America*, 110(5), 2687-2687.

Houston, D.M., Jusczyk, P., Kuijpers, C., Coolen, R. & Cutler, A. (2000), Cross-language word segmentation by 9-month olds, *Psychonomic Bulletin and Review*, 7, 504-509.

Houston, D.M., Santelmann, L.M. & Juszcyk, P. (2004), English-learning infants' segmentation of trisyllabic words from fluent speech, *Language and Cognitive Processes*, 19, 97-136.

Jones, M.R. (1987), Dynamic pattern structure in music: recent theory and research, *Perception and Psychophysics*, 41, 621-634.

Jusczyk, P., Luce, P.A. & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language, *Journal of Memory and Language*, 33, 630-645.

Jusczyk, P., Houston, D. & Newsome, M. (1999), The beginnings of word segmentation in English-learning infants, *Cognitive Psychology*, 39, 159-207.

Ladd, D.R. (2008), *Intonational phonology* (2nd edition), Cambridge: Cambridge University Press.

Ladefoged, P. (1975), *A course in phonetics*, New York: Harcourt Brace Jovanovich.

Lee, C.S. & Todd, N.P.M. (2004), Towards an auditory account of speech rhythm: application of a model of the auditory 'primal sketch' to two multi-language corpora, *Cognition*, 93, 225-254.

Lerdahl, F. & Jackendoff, R. (1983), *A generative theory of tonal music*, Cambridge, MA: MIT Press.

Low, E., Grabe, E. & Nolan, F. (2000), Quantitative characterisation of speech rhythm: 'syllable-timing' in Singapore English, *Language and Speech*, 43 (1), 377-401.

Nazzi, T., Bertoncini, J. & Mehler, J. (1998), Language discrimination by newborns: towards an understanding of the role of rhythm, *Journal of Experimental Psychology: Human Perception and Performance*, 24, 756-766.

Nazzi, T., Iakimova, G., Bertoncini, J., Fredonie, S. & Alcantara, C. (2006), Early segmentation of fluent speech by infants acquiring French: emerging evidence for cross-linguistic differences, *Journal of Memory and Language*, 54, 283-299.

Otake, T., Hatano, G., Cutler, A. & Mehler, J. (1993), Mora or syllable? Speech segmentation in Japanese, *Journal of Memory and Language*, 32, 358-378.

Parncutt, R. & Drake, C. (2001), Psychology: rhythm, in *New Grove dictionary of music and musicians*, Vol.20 (S. Sadie, editor), 535-538, 542-553.

Patel, A.D. (2008), *Music, Language, and the Brain*, Oxford: Oxford University Press.

Pike, K. (1945), *The intonation of American English*, Ann Arbor: University of Michigan Press.

Plack, C.J. & Carlyon R. (1995), Loudness perception and intensity coding, in *Hearing: handbook of perception and cognition*, 2nd edition (B.C.J. Moore, editor), San Diego: Academic Press, 122-160.

Ramus, F. & Mehler, J. (1999), Language identification with suprasegmental cues: a study based on speech resynthesis, *Journal of the Acoustical Society of America*, 105(1), 512-521.

Ramus, F., Nespor, M. & Mehler, J. (1999), Correlates of linguistic rhythm in the speech signal, *Cognition*, 73, 265-292.

Saffran, J.R., Aslin, R.N. & Newport, E.L. (1996), Statistical learning by 8-month-old infants, *Science*, 274, 1926-1928.

Saffran, J.R., Werker, J.F. & Werner, L.A. (2006), The infant's auditory world: hearing, speech, and the beginnings of language, in *Handbook of child development* (R. Siegler & D. Kuhn, editors), New York: Wiley, 58-108.

Thiessen, E.D. & Saffran, J.L. (2007), Learning to learn: infants' acquisition of stress-based strategies for word segmentation, *Language Learning and Development*, 3(1), 73-100.

Todd, N.P.M. (1994), The auditory 'primal sketch': a multiscale model of rhythmic grouping, *Journal of New Music Research*, 23, 25-70.

Todd, N.P.M. & Brown G.J. (1996), Visualisation of rhythm, time and metre, *Artificial Intelligence Review*, 10, 253-273.

Todd, N.P.M., Lee, C.S. & O'Boyle, D.J. (1999), A sensory-motor theory of rhythm, time perception and beat induction, *Journal of New Music Research*, 28, 5-29.

Todd, N.P.M., O'Boyle, D.J. & Lee, C.S. (2002), A sensorimotor theory of beat induction and temporal tracking, *Psychological Research*, 66, 26-39.

Vroomen, J., Zon, M. van & Gelder, B. Van (1996), Cues to speech segmentation: evidence from juncture misperception and word spotting, *Memory and Cognition*, 24, 744-755.

White, L. & Mattys, S.L. (2007), Calibrating rhythm: first language and second language studies, *Journal of Phonetics*, 35, 501-522.

APPENDIX – THE LEE/TODD (LT) SENTENCES

The **count notes** that his **debts** are im**mense**.
Le **comte note** que ses **dettes** sont im**menses**.
Je**rome** and Ma**rie** have **caught** the **bus**.
Jé**rôme** et Ma**rie** ont **pris** le **bus**.
The **wine** is **dear** in the **bars** he fre**quents**.
Le **vin** est **cher** dans les **bars** qu'il fré**quente**.
The **sketch** will **please** both Ma**ri**a and **Paul**.
L'es**quisse** va **plaire** à Ma**rie** et à **Paul**.
The **Turk** ob**served** his **wife** near the **park**.
Le **turc** ob**serve** sa **femme** près du **parc**.

Stressed syllables are highlighted.

SPEECH RHYTHM AND TIMING: STRUCTURAL PROPERTIES AND ACOUSTIC CORRELATES

Antonio Romano
Laboratorio di Fonetica Sperimentale 'Arturo Genre'
Dipartimento di Scienze del Linguaggio – Università degli Studi di Torino
antonio.romano@unito.it

1. ABSTRACT

In my intervention to the Round Table I summarised results from a selection of recent contributions to the research on rhythm and speech timing coming from two Italian laboratories: the *Laboratorio di Linguistica* of the *Scuola Normale Superiore di Pisa* and the *Laboratorio di Fonetica Sperimentale 'Arturo Genre'* of the University of Turin.

In my short presentation I emphasised reference to papers by Pier Marco Bertinetto and Chiara Bertini (Bertinetto & Bertini, 2008, forthcoming; Bertini & Bertinetto, 2009) and Paolo Mairano and Antonio Romano (Mairano & Romano, 2007, 2010), with an introduction explaining the reasons of my own interests in it.

2. PERSONAL INTERESTS AND MOTIVATIONS

Since the early '90s, even though mainly working on intonation structures for my *PhD* (Romano, 1999, 2001), I reserved a relevant interest in speech timing (not neglecting references to other connected fields) and I have been studying the basic literature on this topic (early intuitions of Pike and Abercrombie, more specific contributions by Allen, Bertinetto, Dauer, Fowler, Lindblom, Miller, Roach and many others) oriented to the description of diverging tendencies shown by languages in terms of speech timing and rhythm in production and perception.

Since my research focused on the prosodic properties of Italian dialects, in an early description of the suprasegmental properties of Southern Italian varieties I made an attempt to include considerations on rhythm, following suggestions from Bertinetto (1977), Dauer (1983), Bertinetto (1989) and Schmid (1996).[1] For that research I had the opportunity to study a considerable amount of literature that had appeared in the early '90s on timing and rhythm of Italian dialects (which is now sometimes neglected by scholars working on Southern Italian varieties).[2]

[1] Sallentinian and Apulian dialects are well told apart on the basis of rhythmic properties (a description is now in Romano 2003). In Molinu & Romano (1999) we also took into account measures and experimental work on syllables and other relevant structural properties of some of these dialects.

[2] For instance, many papers published by John Trumper's research team (see, e.g., Mendicino & Romito, 1991, and Romito & Trumper, 1993) refer to evaluation methods proposed by Lindblom & Rapp (1973). Results as well as other research indications suggested by these instrumental works pointed out significant differences in this linguistic domain. As proved by Schmid (2004, 2008), Italian varieties (even within the same broad area, such as

However, all my studies on this topic were carried out before the proliferation of new analysis techniques after the proposal by Ramus, Nespor & Mehler (1999). Research on rhythm exploiting this new approach started in our laboratory in more recent years, when I had the chance to meet Paolo Mairano, who helped me to rapidly get an insight on how rhythm research had evolved in the last decade, and when we found similar interests with the staff of the Laboratory of Linguistics in Pisa.

3. THE RELATIONSHIP BETWEEN TIMING AND RHYTHMIC PATTERNS

The very topic of my presentation started with a rapid hint to the relationship that ties timing and stress- or syllable- patterns. After a short discussion on terminological issues, the terms of a scientific divide were reviewed.

3.1 Terminological issues and cause-effect doubts

We know that the label 'stress-timed' is reserved to languages whose timing is dominated by stress patterns and that the corresponding label 'syllable-timed' refers to languages whose timing is regulated by segmental time patterns depending on syllabic constraints.

A natural typology of languages where foot vs. syllable seems to dominate the metric organisation of speech is confirmed by the observation of different versification traditions in the world's languages and by their variable predisposition to fit in musical-rhythmic frames (cp. Pamies, 1999, referring to Sachs, 1953), but is mainly based on linguists' intuitions which are often influenced by impressions.[3] It has been claimed that languages of the former type exhibit isochrony at the foot level, while languages of the latter type are said to exhibit isochrony at the syllable level (Abercrombie, 1967; cp. Crystal, 1994).

In fact, these alleged isochronies have not been verified in speech and, as it has been proved by several authors (cp. Roach, 1982), languages do not show this kind of metrical regularity when we observe them in connected speech. This way, the isoaccentual/ isosyllabic divide we refer to, in Italian for instance, seems to have little to do with real phonetic rhythmic cues – 'isochrony' is used by linguists and dialectologists to account for vocalic lengthening patterns and to explain distribution rules observed in some dialects.[4]

Furthermore, once that metrical regularities are not confirmed in connected speech, even the stress-/syllable-time classes should be reanalysed in broader terms (as proposed by Stephan Schmid during this workshop), namely distinguishing *stress-based* (STB)

---

mid-southern ones, represented in his data by Neapolitan and Bitontino) show rather different rhythmic patterns and may not be accounted for without due distinctions.

[3] The original distinction emerges from auditory cues. It is well stated everywhere that we owe these definitions to Lloyd James (1940: 25), distinguishing "machine-gun" vs. "morse-code" languages, and to Pike (1945: 35) who refers to syllable-timed vs. stress-timed languages.

[4] As for other less used labels, such as 'isochronicity', perhaps we have a more useful candidate in order to illustrate this structural phonetic regularity on acoustic grounds. A critical view of these language properties, based on instrumental research carried out on spontaneous speech as well as on pre-planned speech, is now in Giordano (2008), also reviewing various sources not confirming such regularities in Italian varieties.

vs. *syllable-based* (SYB) languages independently of stress or syllable phonological constraints which could be compensated at a phonetic level (see the examples of Arabic evaluated by Ghazali *et al.*, 2002; see §5.1).[5]

Several questions persist, however, which cannot be answered merely by a terminological discussion. We still do not know whether there are different timing models or a unique model with local preferences. We still do not know whether rhythm does emerge from other structural properties or rather is a primitive linguistic variable (even unconsciously) controlled in production.[6] In other terms, as discussed by Krull & Engstrand (2003), we do not even know if rhythm is a phonological variable or the phonetic consequence of other phonological events.

3.2 Other issues related to scientific divides

Along this line, I pointed out another issue whose evidence comes again from personal notes. I carried out my *PhD* research in Grenoble (France) in the years when research on rhythm was gaining ground (thanks to Ramus' *PhD*) but I attended a series of conferences on timing where rhythm measuring was not mentioned, probably because it was considered to pertain to different research scopes.[7] What's more, this happened in a laboratory next to the one where Plinio Barbosa was preparing his own *PhD*:[8] the Ramus' approach was not considered a key contribution by the research team with whom Barbosa was working with and we may say that, somehow, it is still considered at least not directly related to linguistic rhythm modelling.

The two main approaches to speech rhythm (modelling and measuring) are still kept separated, as is also proved by connections to research on speech perception and production. That could be shown by another anecdote. In 1995 I was appointed research assistant in the Research Centre where Kate Demuth was working to her *bootstrapping* model whose main account was to be published in the following years (Demuth, 1996), without a reference to rhythm evaluation methods used in Europe at that time (in laboratories of the same town). Yet, in the following years the interest for rhythm in speech perception and language acquisition awoke newly and many results were published by several

---

[5] See the discussion in Vayra *et al.* (1984) and references below (§2.3). Perception issues are raised by Allen (1975).

[6] Generally speaking, rhythm is explicitly mentioned as an organising principle of linguistic timing by Lenneberg (1967). Rhythmical regularities present in the phonology of languages have been extensively studied in terms of prominence patterns by various scholars (see e.g. Liberman & Prince, 1977; Hayes, 1984; Nespor, 1993).

[7] 'Timing' may be referred to co-articulatory properties between segments in intrasyllabic environments as well as to internal durational properties of syllabic or rhythmic patterns, or even to temporal relations between units in a macro-rhythmic environment (feet or prosodic words, sentences...). Research made within intra-/inter-syllabic environments is summarised in works by Rudolph Sock (who was an invited speaker of one of these conferences; see, e.g., Sock *et al.*, 1996).

[8] Furthermore, my *PhD* supervisor, Michel Contini, was in the dissertation panel of Barbosa's *PhD*.

scholars, including Demuth, showing rhythm as a base for *bootstrapping*[9]. The relevance of rhythm in language acquisition has been proved at least since Mehler *et al.* (1981) and Miller (1984), while several studies of the '90s (see e.g. Nazzi, Bertoncini & Mehler, 1998) added experimental evidence. And again, in the same years, MacNeilage's theory 'frame, then content' was brought to completeness, leading to an articulatory acknowledgement of the role of rhythm in speech production and language acquisition (mainly in relation to mandibular oscillations; cp. Rhardisse & Abry, 1995) without explicit mutual connections between these studies (see MacNeilage & Davis, 1990, MacNeilage, 1998).

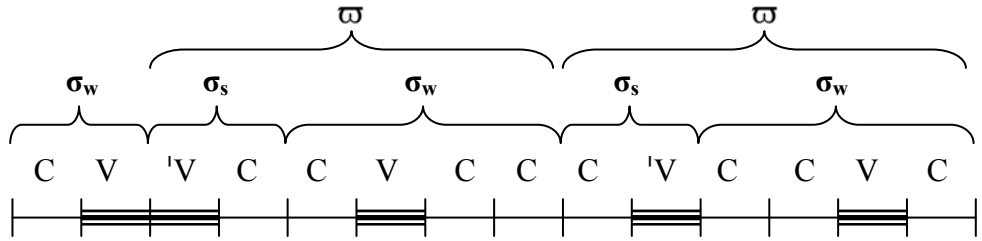


Figure 1: Foot analysis of a fictitious weak-strong syllable sequence and different syllable types

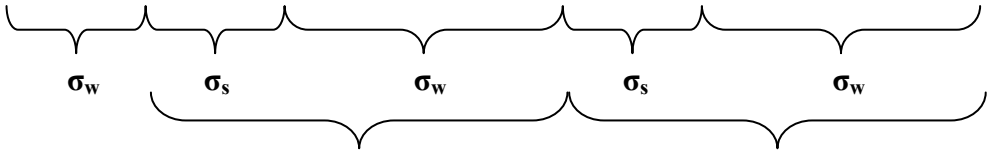


Figure 2: Units which would be taken in consideration for the fictitious utterance in Figure 1 by a phonological evaluation – durations of syllables and feet (interstress patterns)

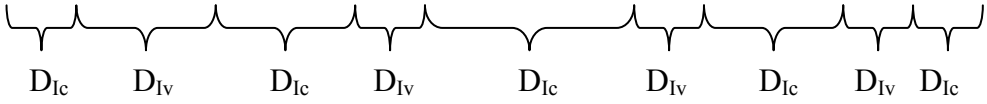


Figure 3: Units which are taken in consideration for the fictitious utterance in Figure 1 by recent rhythmic measures – durations of vocalic and consonantal intervals.

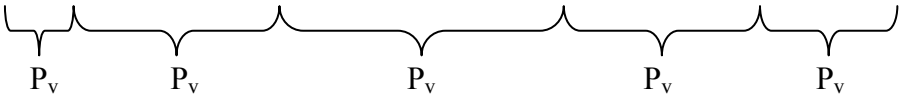


Figure 4: Units which are taken in consideration for the fictitious utterance in Figure 1 by traditional rhythmic measures – durations of intervals between vocalic pulses

[9] Within the framework of generative approaches, one may find interesting phonetic data in Frota, Vigário & Martins (2002) and Frota, Vigário & Freitas (2003). For a general reference see Fikkert (2007).

This shows as, in those years, significantly different approaches to the study of time variables in speech focused on rhythm modelling and timing evaluation (even though they did not always found a joint interest in it for their respective research fields).

3.3 Different approaches

As it has been said, the history of speech rhythm research (summarised in the bibliography collected by P. Roach) could be divided in a *before* and an *after*. In the *before*, to which authors sometimes go back for new proposals aiming at testing more robust models, we found experimental studies on the resistance of syllable- or interstress-patterns to the compression.

In these studies, measures were carried out at the syllable level, by evaluating foot or syllable durations and their resistance to tempo variation or with increasing lengths of utterances (Figure 2).[10]

Despite this dominant approach, at that time, there were already studies pointing out the relevance of vocalic pulses in rhythm perception (see Allen, 1975; Barbosa, 2006; also cp. Miller, 1984). They proposed measures as in the third method here recalled (based on V-to-V intervals, related to the so-called *P-centre* method, Figure 4).

In Figure 3, we show instead measures suggesting Consonantal and Vocalic Intervals as relevant units, following Ramus and colleagues proposals which determined the starting point of the *after*.

4. SPEECH RHYTHM MEASUREMENTS BASED ON DURATION

Within the framework defined by Ramus and colleagues, Mairano & Romano gave a few early contributions in rhythmic typology since 2006 – and that, without developing new models or theories on speech rhythm. Nevertheless, we consider them to be fairly original because they raise relevant questions.

We started by measuring stretches of V and C as other colleagues were doing in different European laboratories and, in the same period, we shared with Pier Marco Bertinetto our impressions about the different rhythmic metrics and doubts about which kind of speech make-up to use in order to test them. We observed a lack of consistency in certain basic assumptions made by rhythm specialists and the need to be more explicit on working hypotheses often implicit in recent papers, namely the specific choices made in the segmentation of speech files and during the statistical processing. Among sources of variability in the estimation of rhythm metrics, we investigated the agreement between different operators in the classification of V-C items.

[10] We may refer to, e.g., Marotta (1985), who studied sentences like: *Perciò pésa((me)lo) tutto di nuovo... Perciò pesate((me)lo) tutto di nuovo...* Similar sentences were measured and tested by other authors among which I would like to mention A. Pamies (see references in Pamies, 1999) who brought evidence on the reduced compression properties of Spanish and French and discussed the discriminating role of stiffness parameters related to syllable and segment durations. Similar outcomes are discussed for Italian by Bertinetto, (1977, 1983) and Farnetani & Kori (1986, 1990).

In particular, in Mairano & Romano (2007a, 2007b, 2008a):

1) we wondered whether the classification of segmental units should be done phonologically or phonetically and we raised questions about the classification of velarised and vocalised laterals, rhotic elements in coda, syllabic sonorants and so on;

2) we proposed to test the sensitivity of the metrics to segmentation choices made by different operators (discussing at the same time problems concerning speaker variation and speech rate conditions);

3) we observed different methods for determining the metrics and we tested the differences in the results obtained joining together the metrics from different interpausal units or, instead, keeping separate statistics for them.

We showed how these elements influence the final values of the metrics and introduce relevant changes in the rhythmic topology of languages. Figure 5 shows an example relating to the first issue (see Mairano & Romano, 2007a). In this example (from German) the syllabic sonorant (the second nasal segment) has been analysed as vocalic even though it has less energy than the other two surrounding nasals which are told apart as consonants on a phonological basis. Thus, we classified the syllabic sonorant in *seinen Mantel* as a V. But how many automatic tools used for segmentation would have detected it and classified it like that?

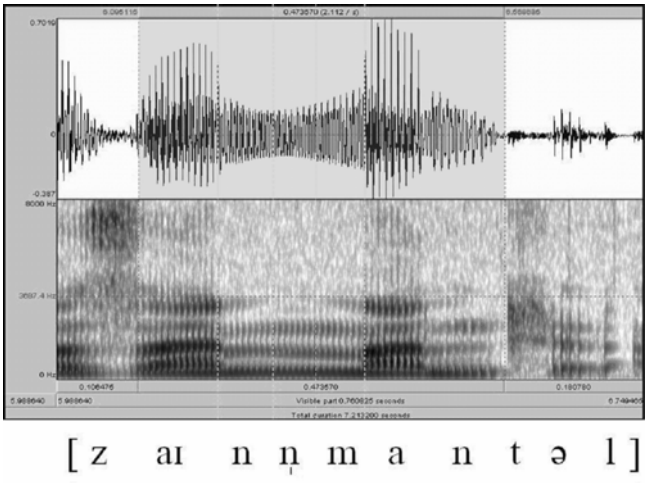


Figure 5: Which segmental cues justify the classification of a segment as V or C?

Another relevant question we raised is then: should rhythm be evaluated in terms of expectations or of real facts?

Still working on the traditional typology which classifies languages along the stress-/syllable-timed axis, we tried to assess how much the results might depend on operators' segmentation and classification choices during measurement tasks (see Figure 6; cp. Mairano & Romano, 2007a & b). We confirmed a specific, sample-dependent, language conditioning and we discussed this topic at the same time. However, we enlarged the scope of

similar arguments in a poster presented at the last *ICPhS* in Saarbrücken (see Mairano & Romano, 2007b).

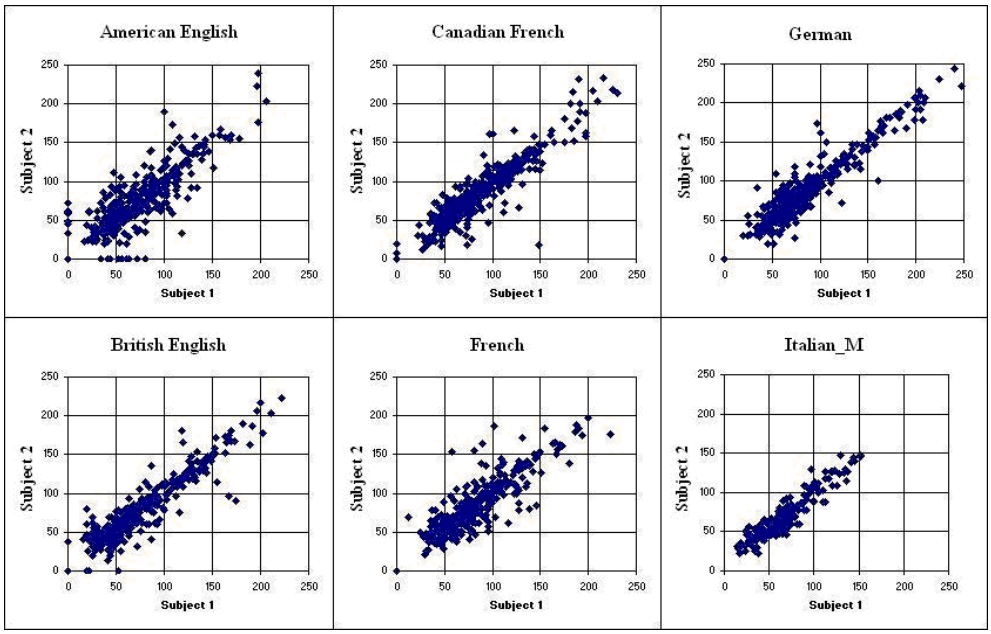


Figure 6: How much do operators agree in segment classification and measurement?

In these plots (from different language samples) two operators separately judged the duration of intervals. Values are variously scattered (sometimes significantly) from the regression line. A few segments have been measured as consonants by one operator (especially /r/ realisations in American English) whereas the other one considered them as belonging to rhotacised and diphthongised vocalic nuclei (duration as C=0).

An aspect on which authors are not explicit when presenting their results is how the metrics are computed, whether on the entire production (which we call A-mode) or by averaging partial results on each interpausal interval (B-mode). The question is particular relevant when dealing with speech samples coming from spontaneous dialogues which are naturally segmented at least by turn-taking and because the two speech production lines are intertwined.

We tested the effects of the two ways of computing metrics in our speech make-up (continuous monologue productions) and we showed significant changes (see Figure 7 for a comparison between the deltas computed for the same six samples: the rhythmic topology of languages changes when switching from the A-mode (on the left) to the B-mode (on the right).[11]

[11] Plots are made with the software *Correlatore* (by P. Mairano, available at http://www.lfsag.unito.it/correlatore/index_en.html).

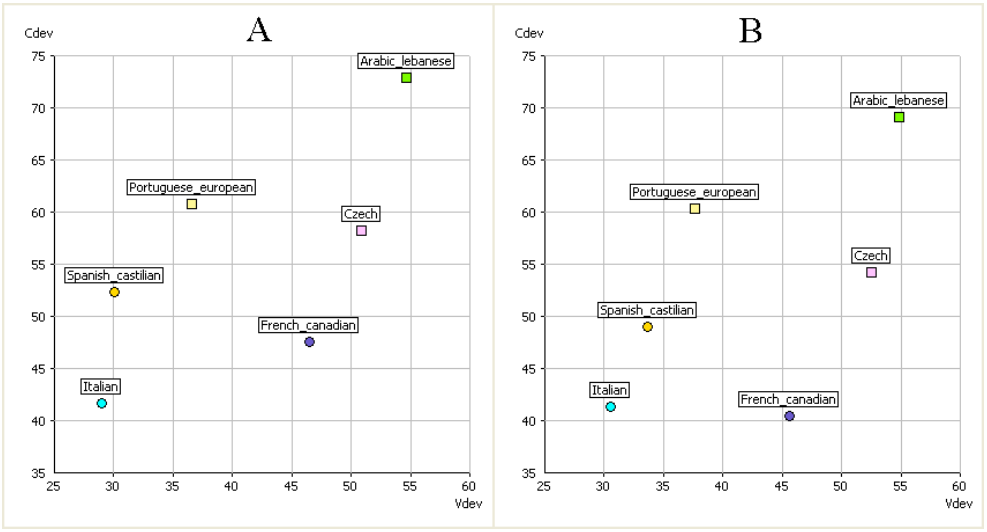


Figure 7: Effects of joint (A) vs. separate (B) statistic computations for the metrics proposed by Ramus *et al.* (1999)

Differences have been tested also on samples uttered by different speakers for the same language. Sensible differences are detectable for languages such as Italian, for speakers from different regions, even when dealing with standard-like samples and have been confirmed for ten speakers of Icelandic (which has virtually no geographic variation at all) showing a fairly strong dependence on speaking styles.

The discussion about this point was hinted at in our previous papers, but final results were discussed at the EASR 2008 workshop in UCL (Mairano & Romano, unpublished, poster presentation), together with the issue about which spoken corpora should be used.

Further discussions took place in Granada during the presentation of Russo & Barry (2008a) about the kind of speech make-up on which to test models. The questions are:

1) Are we able to perceive linguistic rhythm by listening to the speech of our interlocutor?

2) If yes – as it must be, since we began discussing about this topic from Pike's and other linguists' intuitions –, in what kind of linguistic productions are we able to detect enough cues to classify the linguistic rhythm of our interlocutor and of his/her language?

We got fairly good results by computing the simple metrics proposed by Ramus *et al.* (1999) on the short narratives offered by the *IPA Illustrations* in which there are enough rhythmic cues in order to allow a trained listener to guess a rhythmic classification. 'Fairly good results' means that the positioning of languages in a continuous space is in accordance with listeners' intuitions.

5. SPEECH RHYTHM MODELLING

Speech rhythm modelling received considerable attention in the last years. In this section of my intervention, I made reference to some recent contributions, accounting for linear regression studies (summarised in Barbosa, 2006) and new multi-layer models (discussed in Bertinetto & Bertini, forthcoming). The duration of the stress group (Interstress Interval) is defined as a function of the number of syllables (*n*). The well known formula is the following one:

$$I(n) = a + b \cdot n \quad (1)$$

where *a* is a constant and *b* is a parameter describing the growing ratio of *I* versus *n*.

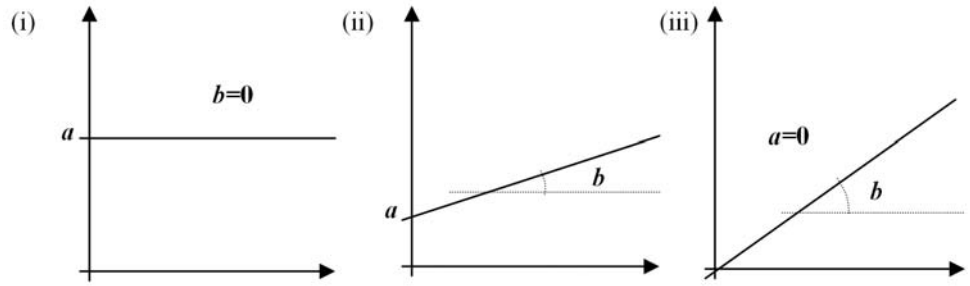


Figure 8: The growth of Interstress Intervals for (i) absolute stress-timed languages (on the left), (iii) for absolute syllable-timed languages (on the right) and (ii) for a mixed-timed language (in the mid)

With this formula, the two extreme ways of establishing the priority in rhythmic regulation of different languages are:

- an absolute stress-timing, when *b* is naught and, therefore, the Interstress Interval is a constant ($b=0 \rightarrow I=a$; see Figure 8 (i));
- an absolute syllable-timing, when *a* is naught and the Interstress Interval is directly proportional to the numer of syllables ($a=0 \rightarrow I=bn$; see Figure 8 (iii));
- but languages usually tend to show an intermediate way (see Figure 8 (ii)).

5.1 The double oscillator model

For long-term qualitative descriptions of language timing, a model has been re-proposed – see the relevant literature on previous studies, e.g. in Barbosa (2006) – who predicts temporal patterns as the result of the coupling of two oscillators (see O'Dell & Nieminen, 1999).

The duration of the Interstress Interval is described as the function of the number of syllables and of two clocks whose contributions are regulated by a coupling strength (called *r*-parameter), so that *a*, *b* and *I* of the preceding equation are re-defined as in the following formula (where ω_1 is the oscillation velocity of the accentual oscillator, ω_2 is the oscillation velocity of the syllabic oscillator and *r* is the coupling strength):

$$I(n) = \frac{r}{r\omega_1 + \omega_2} + \frac{1}{r\omega_1 + \omega_2} n$$

(2)

When the value of the coupling strength (r) is 1, then the a of the original equation is equal to b and both oscillators have the same influence; but when r is greater than 1 ($r > 1$) the overarching accentual-oscillator is dominant whereas when r is lesser than 1 ($r < 1$) it is the subordinated syllabic-oscillator which is dominant.

Studies of the '80s-'90s carried out for Swedish and English (Fant, Eriksson and others cited by Barbosa, 2006) have evaluated r on different corpora with changing tempos and have assessed values around 2 against typical values obtained for Italian or Greek ($r \approx 0.9$).

Barbosa (2006) tested the same mathematical model for different speech rates for Brazilian Portuguese finding values about 1.5. But for increasing speech rates r did not systematically decrease, thus not confirming the prediction of more syllable-timed behaviours of the same language for rapid tempos (see Dellwo & Wagner, 2003, for different results).

5.2 The Control-Compensation model

The Control-Compensation model (*CC*) proposed by Pier Marco Bertinetto and Chiara Bertini (2008, 2009 and forthcoming) finds its origins in earlier studies of '80s (see e.g. Bertinetto & Vékás, 1991) and reintroduces the double oscillator model in view of providing a unified, fully explicit and more predictive theory.

The model is grounded on the reformulation of language differences observed in terms of reduction of vowels (V) and consonants (C) in a gestural overlap hypothesis framework. That leads to a revisited dichotomy not involving the stress-/syllable-timing axis but contrasting more controlling languages (*CTRL*) vs. more compensating languages (*CMPS*).

The rhythmic metrics proposed aim at accounting for intrasyllabic durational stability vs. compression in a dynamic model inspired by the *PVI* model (see Grabe and Low, 2002). Therefore the *CCI* model introduces the number of segments (n) in the metrics.

The formulae defining the two indexes *CCI(V)* and *CCI(C)* are:

$$CCI(V) = \frac{100}{n_{I_V} - 1} \cdot \sum_{k=1}^{n_{I_V}-1} \left| \frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}} \right| \quad \text{and} \quad CCI(C) = \frac{100}{n_{I_C} - 1} \cdot \sum_{k=1}^{n_{I_C}-1} \left| \frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}} \right|$$

(4) (5)

with n_{IV} and n_{IC} the numbers of Vocalic and Consonantal Intervals in the speech sample and n_k the number of segments in the k interval.

The CCIs are applied to two levels of organisation: a phonotactic level, called 'Level-I', which is based on the coupling of the vocalic and consonantal oscillators (as it is also suggested by Goldstein *et al.*, 2007); a phrasal level, called 'Level-II', which is based on the coupling of the accentual and syllabic oscillators (see O'Dell & Nieminen, 1999).

The formula defining the Interstress Interval was applied to the Level-I oscillator relating the duration of inter-V-onset intervals – from one V-onset to the next (as suggested by Keller & Port, 2007; Barbosa, 2006) – to the number of intervening consonants. The Intervocalic Interval is regulated by r_1.

When r_I is greater than 1 the vocalic oscillator prevails on the consonantal oscillator. This allows to predict that:

- the consonantal oscillator should emerge as dominant along with tempo increases, for the consonants comprised between two vocalic gestures cannot be compressed beyond a certain threshold, whereas vowels allow for more compression;
- in *CTRL* languages, however, due to the relative incompressibility of unstressed Vs, the vocalic oscillator should partly compensate this effect.

These predictions have all been confirmed, at present, in the simulation of Bertinetto & Bertini (forthcoming), even though with a number of adapted considerations for each sample analysed.

Nevertheless, the groupings in rhythmic classes have been reanalysed in terms of an interplay of Level-I and Level-II leading to four ideal groups and results are encouraging (see Table 1 and Bertinetto & Bertini, forthcoming, for a detailed discussion).

TYPE	*LEVEL-I*	*LEVEL-II*	*EXAMPLE*
1	*CTRL*	*CTRL*	*Italian*: relatively simple phonotactics, fairly rigid word stress pattern
2	*CMPS*	*CMPS*	*English*: fairly complex phonotactics, fairly mobile word stress pattern, density of secondary stresses yielding further prominence sites
3	*CMPS*	*CTRL*	*Polish*: complex phonotactics, fairly rigid word stress pattern
4	*CTRL*	*CMPS*	*Chinese*: simple phonotactics, uncertain word stress pattern

Table 1: The four ideal groups emerging from the interplay of Level-I and Level-II (adapted from Bertinetto & Bertini, forthcoming)

Furthermore, the model fulfils a number of epistemological requirements and works as a good starting point for future improvements.

6. COMPARISON BETWEEN METRICS

Discussions on topics like the ones in the §4 are leading to our growing interest towards rhythm models. However, the research which has been carried out in Turin was not aimed until now to test models. We started to apply the delta calculations (see Ramus *et al.*, 1999) on duration measurements made on a multi-language sample, in order to verify how metrics based on duration of consonantal and vocalic intervals were good correlates of known or expected rhythmic types.

Those are the grounds on which we hope to have given a significant contribution until now, including in our measurements a relevant sample of languages and discussing how metrics capture language clustering around STB and SYB rhythmic poles.

6.1 The discriminating power of the deltas

We calculated %V, ΔV, ΔC for different translations of the *The North Wind and the Sun* read by several speakers. This choice has been made in order to focus on controlled

samples instead of spontaneous productions. We argue that wild auto-segmentation procedures should be avoided where possible since the gain in time is counterbalanced by a loss of precision. Moreover, we think there is no urge to get huge amounts of data, as simple listening tests suggest that humans need only a few seconds to distinguish between languages belonging to different rhythm classes (even when they do not know how to express that; see Ghazali et al., 2002).

We propose here a selection of 29 speakers (4 for German and Italian, 3 for English, 2 for (varieties) of each of the following languages: Chinese, French, Finnish, Icelandic, Portuguese and Romanian, and only one for Arabic, Czech, Japanese, Russian, Spanish and Turkish; see §5.3 and Appendix for details).

Language	*%V*	*ΔV*	*ΔC*
French_european	49,4	41,04	39,86
Romanian_muntenian	40,4	32,70	41,33
Italian_1	45,5	28,96	41,71
Finnish_2	46,7	48,99	43,10
Chinese_Mandarin	51,2	43,67	43,88
Icelandic_10	45,4	37,87	46,40
Icelandic_4	46,5	38,27	46,40
Italian_2	48,2	42,00	46,81
Finnish_1	48,6	52,90	46,93
Romanian_moldavian	43,4	40,29	47,44
French_canadian	51,2	46,51	47,57
Italian_3	46,3	39,62	48,65
Portuguese_brazilian (SP)	49,2	46,96	50,39
Italian_4	42,2	30,62	52,14
Spanish_castilian	42,0	30,07	52,31
Japanese_fast	46,0	35,87	53,62
English_GA	42,2	43,75	55,36
Japanese_slow	48,0	39,38	55,93
Chinese_Cantonese	46,7	51,36	57,47
Turkish	44,9	37,96	57,50
Russian	38,3	36,23	57,92
Czech	44,9	50,81	58,21
English_RP	40,7	51,82	58,85
English_Aus	40,6	42,71	59,66
Portuguese_european	43,9	36,60	60,77
German_IPA	46,9	47,00	60,79
Zurich_German_	50,0	52,68	65,57
German_1	46,1	53,66	65,93
German_2	42,8	47,30	69,85
Arabic_lebanese	46,8	54,69	72,92

Table 2: Delta values for the 30 language samples analysed (ranked on ΔC basis)

Results are fairly encouraging, allowing a reasonable matching between calculated values and impressionistic expectations. As we show in Table 2, samples rank quite well in two classes (in the upper half SYB languages and in the lower half STB languages; cp. plots Figures 9 and 10).[12] That happens if samples share the same general recordings quality and capture a similar degree of fluency and spontaneity for the represented language.

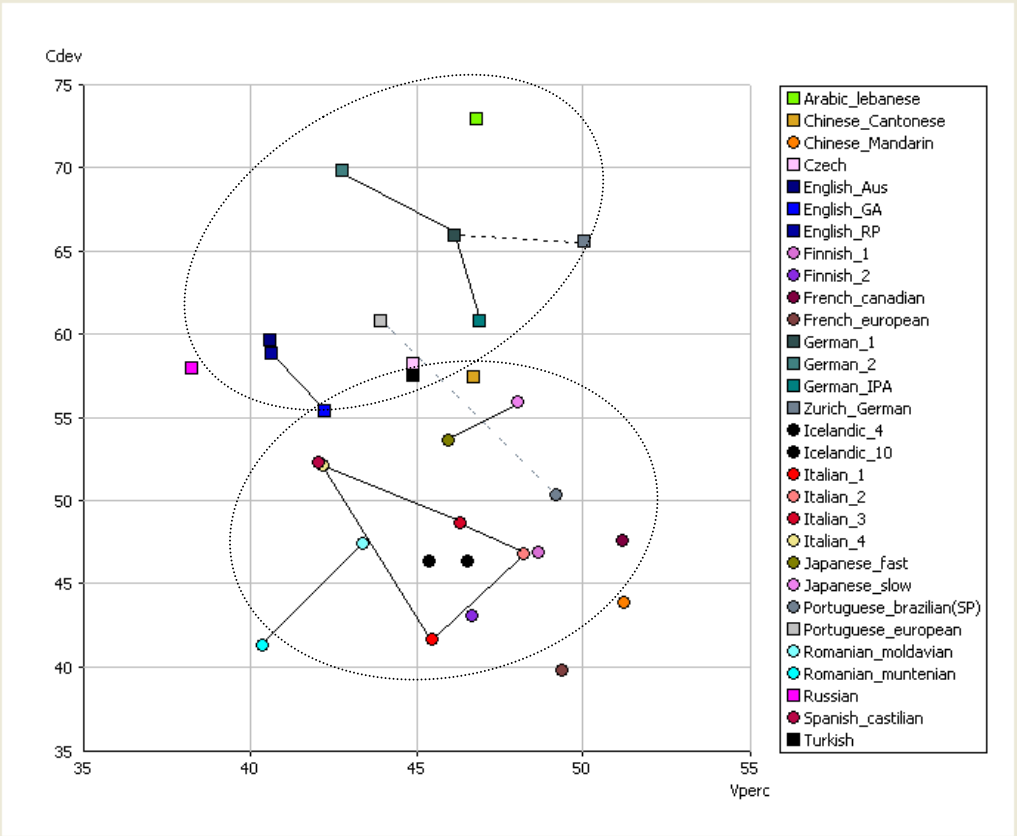


Figure 9: ΔC vs. %V plot for the language samples in Table 2

The metrics calculated for different samples of German (improperly including Zurich German), English, Italian, Portuguese, Romanian and Japanese have been connected in order to allow an easier reading for languages represented by more than one sample.

In some particular cases, one may ask whether the fact the e.g. Arabic ranks at the bottom of this list means that it is a stress-based language. In this respect we believe that the use of such labels per se does not constrain a judgement on the phonological level; only a phonetic indirect evaluation is concerned (further comments in §5.2).

[12] Plots are made with *Correlatore* (by P. Mairano, available at http://www.lfsag.unito.it/correlatore/index_en.html).

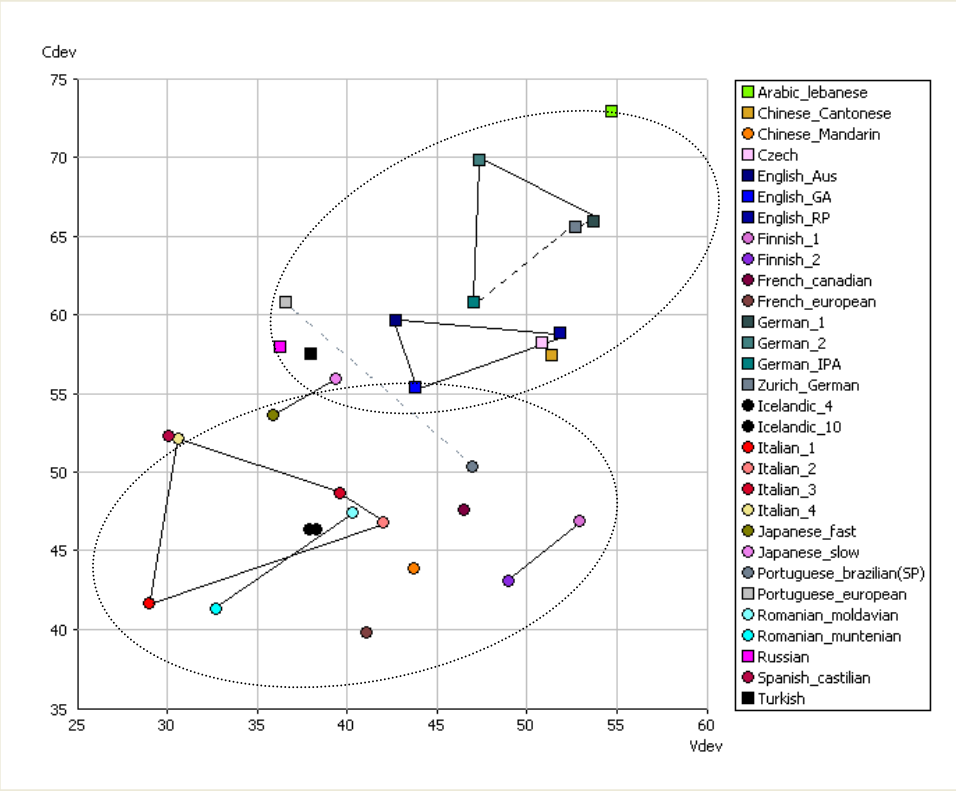


Figure 10: ΔC vs. ΔV plot for the language samples in Table 2[13]

We did not study the effects caused by speech rate (which have been extensively treated in Dellwo & Wagner, 2003, and others) as our samples are fairly homogeneous in this respect (5-6.5 syllables/s) as well as about prosodic segmentation (12-23 silent pauses). The only exception is Japanese which is represented by two samples realised at different speech rates (slower, 3.8 syllables/s, and faster, 4.6 syllables/s). We kept the two samples because – like for other samples we observed before (e.g. Italian in Mairano & Romano, 2007a) –, according to predictions in Dellwo & Wagner (2003), faster speech rates move the results of language metrics towards the SYB pole. In the case of our Japanese samples, which lie in the mid of the transition region along the SYB-STB continuum (against evidence proposed in other studies, e.g. Ramus *et al*., 1999), this parameter determines a critical placement (cp. Barbosa, 2006). See Mairano & Romano (2010) and Figure 11 for a discussion of the sensitivity of metrics to a different treatment of devoiced close vowels in this language.

[13] The metrics calculated for different samples of German, English, Italian, Portuguese, Romanian and Japanese have been connected as in Figure 9.

6.2 Comparison between different metrics - Formulae

Our contributions are restricted to formulae adaptation and metrics comparison for a growing sample of languages (now including Romanian varieties, Czech, Russian, Turkish, Japanese and Chinese, see §5.1) with the definition of a robust evaluation procedure (see above). Like White *et al.* (2007), we believe that – depending on speech rate and style – rhythm metrics may yield rhythmic discrimination of languages along a continuum between the two poles and, perhaps better, in a multidimensional space.

As it is well known (cp. §5.1), Ramus *et al.* (1999) proposed three rhythm metrics (ΔC, ΔV and %V) which are intended to distinguish between two (or four) poles in the space of possible rhythmic organisations (in three charts: ΔC vs. %V, ΔV vs. %V and ΔC vs. ΔV).

Other rhythmic metrics have been proposed in order to decrease the sensitivity of these representation to speech rates and speech styles. The PVIs (*nPVI(V)* and *rPVI(C)* by Grabe & Low, 2002) and the Varcos (*VarcoV* and *VarcoC* by Dellwo & Wagner, 2003) attempt to normalise the effects of speech rate on rhythmic parameters, while the CCIs (controlling and compensating indexes, *CCI(V)* and *CCI(C)* by Bertinetto & Bertini, 2008), a modification of the PVIs, are an attempt to measure the degree of compensation that a language allows for and are inspired by previous work by Fowler (1977) and based on Bertinetto & Vékás (1991; see above §4.2).

We summarise here the formulae for those metrics that have been applied to consonantal and vocalic intervals:[14]

- Deltas (Ramus *et al.*, 1999)

$$\Delta \mathrm{V} = \sqrt{\frac{\sum_{i=1}^{n_V} (D_{V_i} - \overline{D}_V)^2}{n_V - 1}} \qquad \text{and} \qquad \Delta \mathrm{C} = \sqrt{\frac{\sum_{i=1}^{n_C} (D_{C_i} - \overline{D}_C)^2}{n_C - 1}}$$

(6) (7)

- Varcos (Dellwo & Wagner, 2003; Dellwo, 2006)

$$\mathrm{VarcoV} = \frac{\Delta \mathrm{V} \cdot 100}{\overline{D}_V} \qquad \text{and} \qquad \mathrm{VarcoC} = \frac{\Delta \mathrm{C} \cdot 100}{\overline{D}_C}$$

(8) (9)

[14] Dellwo *et al.* (2007) use voiced-unvoiced parameters which seem promising except for the fact that they are prone to show sensitivity to cases of voicing undecidibility and, on a phonological level, they need to be tested on those languages where voice contrasts are not pertinent. Gibbon & Gut (2001) refer to the RIM (Rhythmic Irregularity Measure) defined by Scott *et al.* (1986) and apply their *Rhythm Ratio* (RR, an early modification of the PVI) to syllables and vowels.

- PVIs (Grabe & Low, 2002)

$$nPVI(V) = 100 \cdot \frac{\sum_{k=1}^{n_{I_V}-1} \frac{|d_k - d_{k+1}|}{(d_k + d_{k+1})/2}}{n_{I_V} - 1} \quad \text{and} \quad rPVI(C) = \frac{\sum_{k=1}^{n_{I_C}-1} |d_k - d_{k+1}|}{n_{I_C} - 1}$$

(10) (11)

- CCIs (Bertinetto & Bertini, 2008)

$$CCI(V) = \frac{100}{n_{I_V} - 1} \cdot \sum_{k=1}^{n_{I_V}-1} \left| \frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}} \right| \quad \text{and} \quad CCI(C) = \frac{100}{n_{I_C} - 1} \cdot \sum_{k=1}^{n_{I_C}-1} \left| \frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}} \right|$$

(12) (13)

6.3 Comparison between different metrics - Plots

Following an early evaluation based on deltas, we aimed at testing the variation of results according to a number of factors (as discussed above): we also calculated VarcoV, VarcoC, nPVI(V), rPVI(C), CCI(V) and CCI(C) for the language samples presented in §5.1.

All the metrics yielded to a certain amount of overlapping between rhythm classes and each of them showed sensitivity to slightly different phenomena related to speech timing (cp. Barry & Russo, 2004; Dellwo, 2008). Results are anyway encouraging, confirming our impressions on single samples (even when contrasting with other authors' predictions on rhythm from general features or specific measurements).[15]

In Figure 11 we have an example of a comparison between charts based on different metrics. In particular we may observe how languages are mapped on the CCI map in relation to the bisecting line, along which syllable-timed language are usually placed.

[15] Arbitrarily stating that all Germanic languages are stress-timed (against our impression on Icelandic, see Figure 9), Fikkert *et al.* (2004) surprisingly expect European Portuguese to be a syllable-timed language. The same happens for Czech (our Czech sample sounds more STB, against evidence summarised in Dankovicová & Dellwo, 2007; also see Volín, 2005). Furthermore, our sample of Cantonese lies in the STB area in agreement with the results published by Grabe and Low (2002) (*pace* Mok and Dellwo, 2008; also cp. Jian, 2004). As for the alleged inclusion of Arabic varieties in the STB class see Ghazali, Hamdi & Barkat (2002), even though ΔC gets aberrant values in some tables.

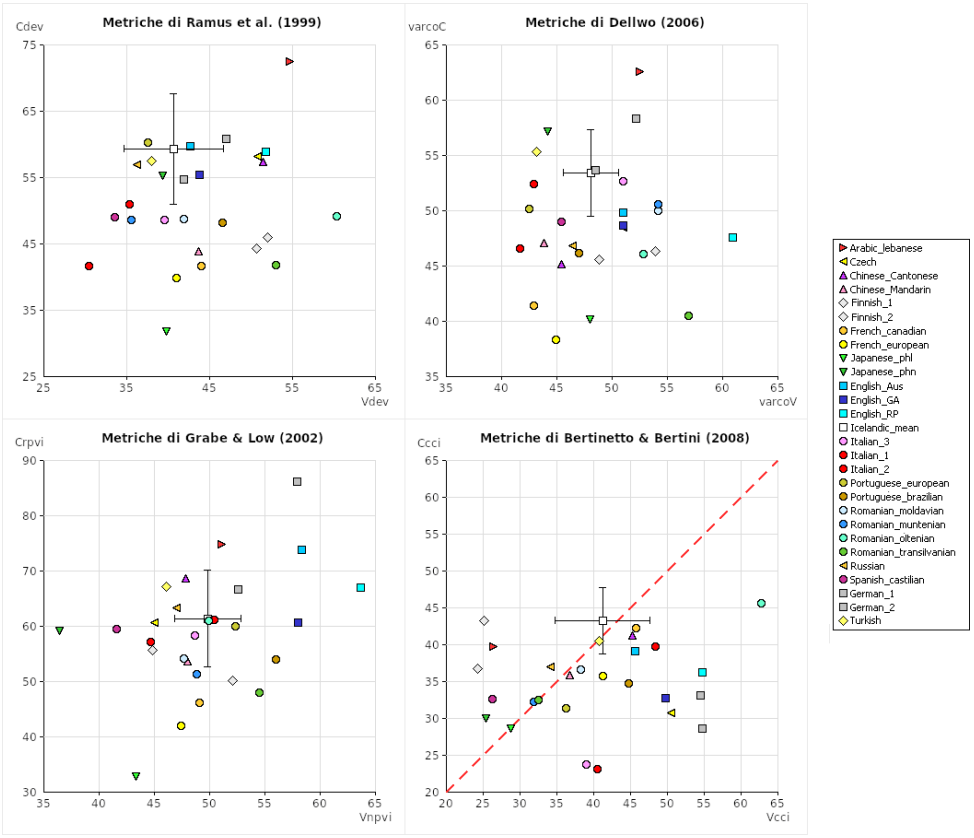


Figure 11: Plots for the same language samples with different metrics
(adapted from Mairano & Romano, 2010)

Dividing the Consonantal and Vocalic Intervals by the number of segments has a correction effect on languages with consonant gemination (pushing a couple of Italian samples towards more STB regions) and on languages with long vowels (occasionally with gemination too; thus dramatically changing the placement of Arabic and Finnish).

As for Italian varieties, we started to analyse data from various dialects. General rhythmic properties of Romance languages are discussed in Mairano & Romano (2009) whereas results on a selection of Piedmontese varieties are presented in Romano, Mairano & Pollifrone (2010).

7. NOT ONLY DURATION

Generally speaking, speech rhythm represents a complex prosodic phenomenon resulting from the co-operative effect of several elements determining strong-weak alternations and it is influenced by various factors (from timing to intonation, within a smaller or a larger scale). Regularity or irregularity in a sequence of pulses may be reflected in sequences of prominences raised by different parameters, among which we mainly expect

to find local variations in the energy, local peaks or movements in the pitch curve and alternating duration patterns (Allen, 1975).

The increasing interest in temporal correlates of rhythm, which seems reasonable above all in comparison with music rhythm, has recently limited the attention of some researchers to these facts and has brought them to neglect other parameters.

As for the questions raised on this topic by the chairman of the Zurich Round Table, Prof. W. Barry, a selection of relevant points have been discussed, among which I chose to answer to the following ones:

- Should rhythm measures be limited to duration? If so, why?
- If not, which other parameters should be included?

In Turin, since the beginnings, we have been wondering whether to include other acoustic correlates in our assessments of speech rhythm (see Mairano & Romano, this volume).

A convincing experiment we have attempted is based on a selection of manipulated sound samples that can be used for a simple informal listening test.

Synthetic stimuli are obtained from two original speech samples using the *AMPER-dat* scripts (for Matlab) which I designed during my *PhD* and which are now adopted within the *AMPER* project testing procedures (see references, *AMPER*).

The two natural speech samples we manipulated here are extracted from the two narratives published for English (RP) and French (Parisian) by the *IPA* (see Roach, 2004, and Fougeron & Smith, 1999). The original utterances are *And at last, the Northwind gave up the attempt* for English (see Figure 12) and *Finalement, elle renonça à le lui faire ôter* for French. The manipulation technique is applied to an already stylised version of the main prosodic content of the sound sample (mainly based on sequences of values for the three parameters, energy, f_0 and duration, for all its vowels).

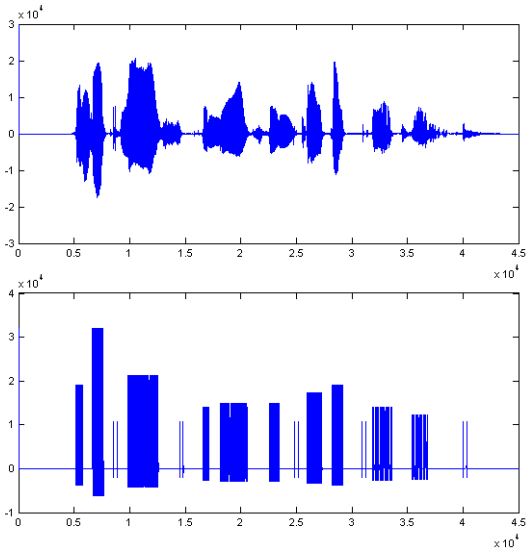


Figure 12: Amplitude plots of the original speech sound sample (up) and the corresponding *.ton* sound file (down)

Figure 12 shows the contrast between the original speech sound sample and the corresponding *.ton* sound file, the latter being a sound sample alternating series of pulses (generated with duration, energy and pitch of each vowel in the original sample) and silences (occasionally broken by isolated pulses representing inner bursts in clusters of obstruents; see Table 3 for an example).

IPA_englishRP_narrative7_F.txt			size: 45725		11-feb-09
	duration [ms]	energy [dB]	fo1	fo2	fo3 [Hz]
1	35	78	263	260	267
2	57	82	301	288	256
3	0	0	50	50	50
4	165	79	185	178	189
5	0	0	50	50	50
6	29	74	204	192	187
7	148	75	174	173	172
8	53	75	195	179	170
9	0	0	50	50	50
10	79	77	171	180	185
11	59	78	208	204	189
12	0	0	50	50	50
13	106	74	144	152	143
14	84	71	154	137	121
15	0	0	50	50	50

values at:
5135 5414 5694 6646 7099 7552 8534 8534 8534 9834 11156 12478 14471 14471 14471 16606 16840 17075 18132 19318 20504 22529 22952 23375 24807 24807 24807 25937 26571 27206 28203 28672 29140 30886 30886 30886 31845 32695 33545 35395 36068 36740 39975 39975 39975

Table 3: An example of text file meant to generate a *.ton* file[16]

Figure 13 shows a comparison between the original prosodic content of a sound sample, in this case the same as in Figure 12 and its stylised version. The amplitude display (up) is associated with the natural course of f_0 (mid) and its close-copy stylisation obtained with the *AMPER-fox* scripts for Matlab (down). Dots on the lower line represent bursts in clusters of obstruents; the plot is based on values of Table 3.

At its origins, this stylisation procedure shares the basic assumptions of the best known close-copy stylisation defined by 't Hart, Collier & Cohen (1990). It provides a synthetic approximation of the natural course of the three prosodic parameters (see Figure 13 for f_0 and duration), meeting two criteria: the prosody of the final sample should be perceptually indistinguishable from the original, and it should be based on the smallest possible number of values stored in a text file (see an example in Table 3).

[16] Segments 3, 5, 9, 12, 15 are isolated pulses, conventionally indicated with 50 Hz values in pitch columns, reproducing internal bursts within C clusters or final (pre-pausal) bursts.

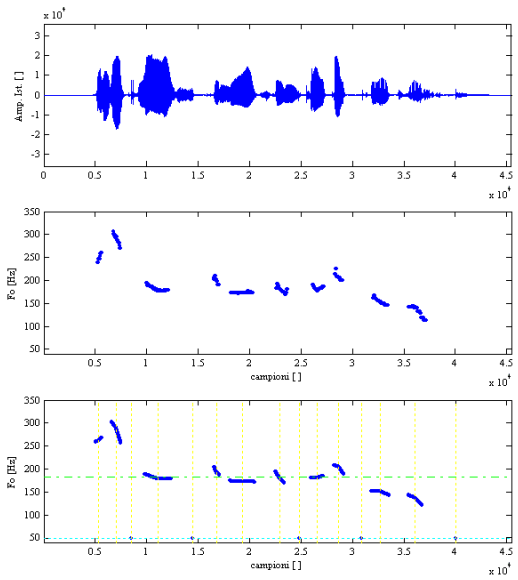


Figure 13: A comparison between the original prosodic content of a sound sample and its stylised version

Our listening tests are based on *.ton* sound files which are series of pulses synthesised from a text file of such kind.

After the two natural sound files for the two samples above (English, supposed STB, and French, supposed SYB),[17] we propose in Figure 14 a selection of manipulated sound files obtained with fixed values for one of the three parameters for all the vowels in the original speech sample.

In Figure 14a. one may listen to the original speech samples and to the synthesised versions with close-copy stylisation of their prosodic content (see the corresponding diagrams for the three parameters D, E, P).

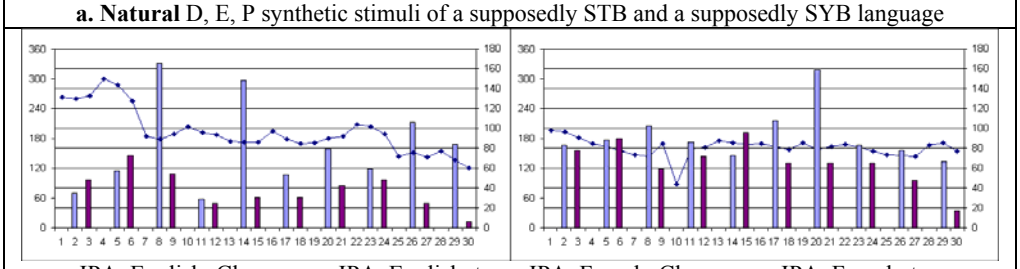
In Figure 14b. one may listen to the *.ton* files with manipulated duration: duration of all the vocalic nuclei in both the original speech samples are fixed to 100 ms (or fixed to the mean value of each sample, which is more natural but keeps the original distinct tempo for both of them). Surprisingly, one notices that, even when the information on the duration of vowels (shown by the higher bars in the diagrams) is lost, the distinct rhythm of the two samples is still present.

Sound files (and diagrams) in Figure 14c. demonstrate the very little contribution of energy in rhythm perception: the lower bars in the diagrams show the normalised E values: their rhythmic characterisation is very similar to that of the original samples.

---

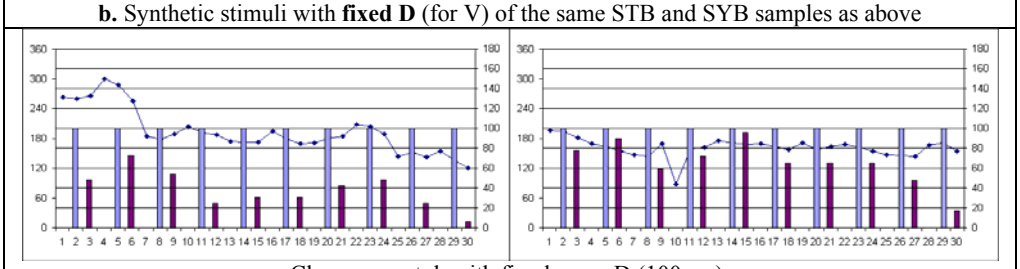
[17] Calculations of deltas on these short samples give about %V=37, ΔV=46 and ΔC=80 for the English utterance (which sounds strongly STB and is therefore well told apart by these metrics) and %V=47, ΔV=18 and ΔC=35 for the French utterance (also prototypically SYB). In this case, we do not care about the languages concerned but simply about the rhythmic properties present in the samples.

a. Natural D, E, P synthetic stimuli of a supposedly STB and a supposedly SYB language



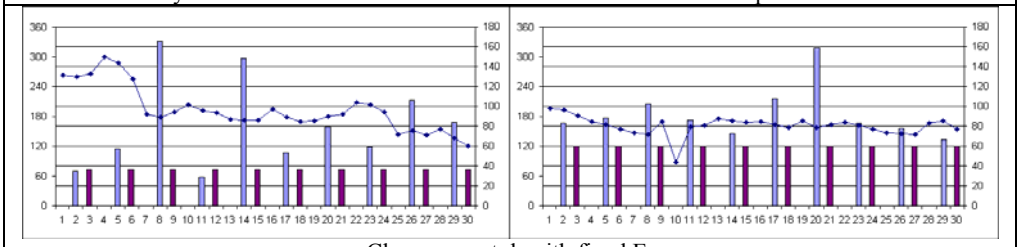
IPA_English; Close-copy: IPA_English_ton – IPA_French; Close-copy: IPA_French_ton

b. Synthetic stimuli with **fixed D** (for V) of the same STB and SYB samples as above



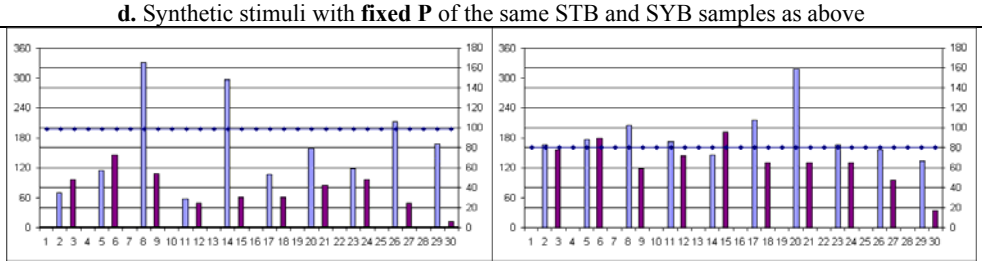
Close-copy styl. with fixed mean D (100 ms):
IPA_English_ton_D100 – IPA_French_ton_D100;
(Close-copy styl. with the original mean D (81 ms & 96 ms):
IPA_English_ton_D81 – IPA_French_ton_D96)

c. Synthetic stimuli with **fixed E** of the same STB and SYB samples as above



Close-copy styl. with fixed E:
IPA_English_ton_E – IPA_French_ton_E

d. Synthetic stimuli with **fixed P** of the same STB and SYB samples as above



Close-copy styl. with fixed P (96 Hz and 80 Hz):
IPA_English_ton_P – IPA_French_ton_P

Figure 14: Curves and sounds for a simple listening test giving an insight into the role of the three different parameters D (duration), E (energy) and P (pitch) in rhythm perception

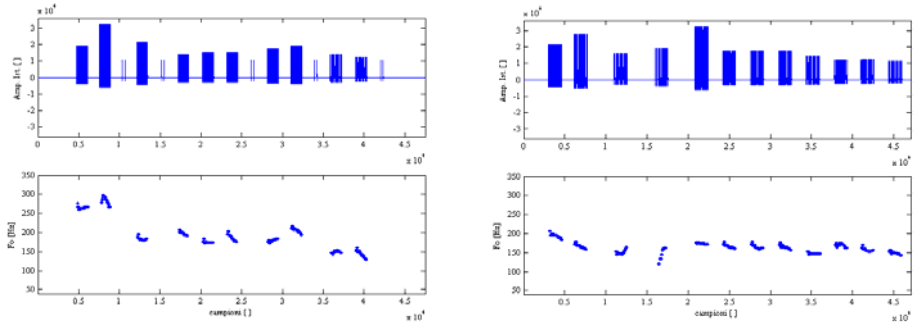
The loss of pitch variations, which can be experience by listening to the synthetic samples in Figure 14d. (with monotonised f_0 respectively at 96 Hz and 80 Hz), seems to prove the relevance of this parameter in rhythmic judgements: the flattening of f_0 information causes a dramatic loss in the different rhythmic characterisation of the two samples.

A simple experiment like this allowed us to start thinking in a different way with regard to speech rhythm:

(1) it demonstrates the inadequacy of metrics based on durations only;

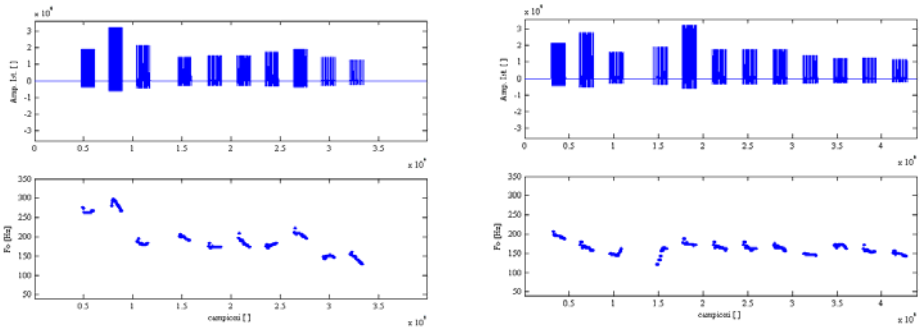
(2) the reduced importance of vocalic durations (observed by listening to the stimuli in Figure 14b.) suggests the possibility that the distance in time between f_0 peaks or specific movements could be one of the main cues in listening discrimination of different rhythmic types (cp. *P-centre* methods).

a. Synthetic stimuli with **fixed D** (for V and C) of the STB and SYB samples as in *Table IV*



Close-copy styl. with the original mean D for V & C (81 ms & 96 ms):
IPA_English_ton_D81_VC – IPA_French_ton_D96_VC

b. Synthetic stimuli with **fixed D** (for V and C clusters) of the two samples (cp. above)



Close-copy styl. with the original mean D for V & C clusters (81 ms & 96 ms):
IPA_English_ton_D81_singleC – IPA_French_ton_D96_singleC

Figure 15: Sounds for a further listening test giving an insight into the role of pitch and duration of V and C.

This hypothesis could be partially invalidated if, even after normalising C durations to a fixed value and thus altering the distances between vocalic pulses, a distinct rhythmic characterisation still persists (listen to the stimuli in Figure 15a.). The possibility of perceptively distinguishing between the two rhythm types is definitely reduced only when the durational information of consonant clusters is limited (listen to stimuli in Figure 15b).

8. CONCLUSIONS

In this intervention I summarised early contributions to rhythm measurement and assessment carried out by the laboratories of Turin and Pisa. I also tried to briefly illustrate the recent attention reserved in Turin to rhythm perception (a listening experimental protocol is being defined by P. Mairano for his *PhD*) in order to include the apparently relevant role of f_0 in speech rhythm evaluation ("otherwise we have a duration model, but not a rhythm model" as stated by Gibbon & Gut, 2001: 96).

Preliminary testing showed that duration could be a good correlate, giving a physical estimate of rhythmic (maybe derived) properties, but we believe that most direct correlates should take into account (distance, extension and shape of) peaks and general melodic profiles.

Linguists' intuitions about speech rhythm are mainly based on perceived or expected properties (often perhaps heavily influenced by the knowledge of phonological features; see e.g. Canepari, 2006, and Ghazali *et al.*, 2002, for examples on how impressionistic evaluations could match experimental/instrumental research).

Even though common people perhaps associate perceived rhythmic properties to stress occurrence conditions or relate it to an impressionistic evaluation of faster vs. slower (or rapid changes in) tempo, we believe that the general intuitions could help to understand the basis of a well assessed dichotomy between two basic rhythm types.

The need to better investigate the relations between stress, intonation and rhythm – which also arises from the simple experiment we proposed in §6 – is also expressed by Giordano (2008). While various authors are addressing their interests in extending the assessment dimensions to other variables (as Lee & McAngus Todd, 2004, with intensity and rhythmograms), other promising results should perhaps derive from different ways to calculates metrics (as it is proposed by Mok & Dellwo, 2008, with DeltaS, or by Gibbon & Gut, 2001, with the rhythm ratio, both accounting for syllable durations).

One of the perspectives for research in this field is to improve rhythm metrics in order to integrate them into a rhythm model (cp. Barbosa, 2006), possibly a multi-layer one, where durational properties could be associated with strong-weak measures or other stress/syllable properties at different levels (as proposed by Bertinetto & Bertini, 2008 and forthcoming; or on the wake of Gibbon & Gut, 2001, distinguishing focal and non-focal components) merge high level (linguistic) information with measures of more than one parameter (namely pitch, duration and perhaps intensity).

ACKNOWLDEGMENTS

I would like to thank Paolo Mairano for the new impulse he has given to our Laboratory, for the analysis tools he realised and for his linguistic help. Many thanks to Carla Marello (University of Turin), Beata Dobrzyńska (CELI), Pier Luigi Salza and Enrico Zovato (Loquendo) for their help in finding reference papers and speakers.

9. REFERENCES

AMPER (*Atlas Multimédia Prosodique de l'Espace Roman*), http://www.lfsag.unito.it/amper.html

Abercrombie, D. (1967), *Elements of General Phonetics*, Edinburgh: Edinburgh University Press.

Allen, G.D. (1975), Speech rhythm: its relation to performance universals and articulatory timing, *Journal of Phonetics*, 3, 75-86.

Barbosa, P.A. (2006), *Incursões em torno do ritmo da fala*, Campinas, Pontes.

Barbosa, P.A. & Albano, E.C. (2004), Brazilian Portuguese, *Journal of the International Phonetic Association*, 34, 227-232 (sound samples available on-line for the *IPA* members: http://web.uvic.ca/ling/resources/ipa/members).

Barry, W. & Russo, M. (2004), Isocronia soggettiva o oggettiva? Relazioni tra tempo articolatorio e quantificazione ritmica, in *Il parlato italiano* (F. Albano Leoni, F. Cutugno, M. Pettorino & R. Savy, editors), Atti del Convegno nazionale di Napoli, 13-15 febbraio, 2003, Napoli: D'Auria (CD-ROM).

Barry, W.J., Andreeva, B., Russo, M., Dimitrova, S. & Kostadinova, T. (2003), Do rhythm measures tell us anything about language type?, in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2693-2696.

Bertinetto, P.M. (1977), *Syllabic Blood* ovvero l'italiano come lingua ad isocronismo sillabico, *Studi di Grammatica Italiana*, 6, 69-96.

Bertinetto, P.M. (1983), Ancora sull'italiano come lingua ad isocronia sillabica, in *Scritti linguistici in onore di G.B. Pellegrini*, II, Pisa: Pacini, 1073-1082.

Bertinetto, P.M. (1989), Reflections on the dichotomy 'stress' vs. 'syllable-timing', *Revue de Phonétique Appliquée*, 91-92-93, 99-130.

Bertinetto, P.M. (1990), Coarticolazione e ritmo nelle lingue naturali, *Rivista Italiana di Acustica*, XIV, 69-74.

Bertinetto, P.M. & Bertini, C. (2008), On modeling the rhythm of natural languages, in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, May 6-9, 2008, 427-430.

Bertinetto, P.M. & Bertini, C. (forthcoming), Towards a unified predictive model of Speech Rhythm, Manuscript.

Bertinetto, P.M. & Magno Caldognetto, E. (1993), Ritmo e intonazione, in *Introduzione all'italiano contemporaneo. Le strutture* (A.A. Sobrero, editor), Roma-Bari: Laterza, 141-192.

Bertinetto, P.M. & Vékás, D. (1991), Controllo vs. compensazione sui due tipi di isocronia, in *L'interfaccia tra fonologia e fonetica* (E. Magno Caldognetto & P. Benincà, editors), Padova: Unipress, 155-162.

Bertini, C. & Bertinetto, P.M. (2009), Prospezioni sulla struttura ritmica dell'italiano basate sul corpus semispontaneo AVIP/API, in *La fonetica sperimentale. Metodo e applicazioni* (L. Romito, V. Galatà & R. Lio, editors), Atti del 4° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Arcavacata di Rende (CS), 3-5 Dicembre 2007, Torriana: EDK Editore, 3-21.

Canepari, L. (2006), *Avviamento alla fonetica*. Torino: Einaudi.

Canepari, L. (2004-), *Fonetica Naturale – Natural Phonetics* (on-line sound samples on Italian pronunciation: http://www.unive.it/canepari [last accessed 30/06/2009]).

Costamagna, L. (2000), *Insegnare e imparare la fonetica*. Torino: Paravia scriptorium.

Crystal, D. (1994), Documenting rhythmical change, in *Studies in general and English phonetics* (J. Windsor Lewis, editor), London: Routledge, 174-179.

Dankovicová, J. & Dellwo, V. (2007), Czech speech rhythm and the rhythm class hypothesis, in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, August 6-10, 2007, 1241-1244.

Dauer, R.M. (1983), Stress-timing and syllable-timing reanalyzed, *Journal of Phonetics,* 11, 51-62.

Dellwo, V. (2008), The role of speech rate in perceiving speech rhythm, in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, May 6-9, 2008, 155-158.

Dellwo, V. & Wagner, P. (2003), Relations between language rhythm and speech rate, in *Proceedings of the 15th International Congress of Phonetics Sciences*, Barcelona, Spain, 471-474.

Dellwo, V., Fourcin, A. & Abberton, E. (2007), Rhythmical classification of languages based on voice parameters, in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, August 6-10, 2007, 1129-1132.

Demuth, K. (1996), The prosodic structure of early words, in *Signal to Syntax*: *Bootstrapping from Speech to Grammar in Early Acquisition* (J. Morgan & K. Demuth, editors), Mahwah, N.J.: Lawrence Erlbaum Associates, 171-184.

Demuth, K. (2003), The status of feet in early acquisition, in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 151-154.

Eriksson, A. (1991), *Aspects of Swedish Speech Rhythm*, Göteborg: University of Göteborg (Monographs in Linguistics, 9).

Farnetani, E., & Kori, Sh. (1986), Effects of Syllable and Word Structure on Segmental Durations in Spoken Italian, *Speech Communication*, 5, 17-34.

Farnetani, E. & Kori, Sh. (1990), Rhythmic Structure in Italian Noun Phrases: A Study on Vowel Durations, *Phonetica,* 47, 50-65.

Fikkert, P. (2007), Acquiring phonology, in *Handbook of phonological theory* (P. de Lacy, editor), Cambridge: Cambridge University Press, 537-554.

Fikkert, P., Freitas, M.J., Grijzenhout, J., Levelt, Cl. & Wauquier S. (2004), Syllabic Markedness, Segmental Markedness, Rhythm and Acquisition, Talk presented at GLOW 2004 (Thessaloniki, Greece, 2004), http://www.del.auth.gr/GLOW2004/abstracts/Fikkert_etal.pdf [last accessed 30/06/2009].

Fleischer J. & Schmid S. (2006), Zurich German, *Journal of the International Phonetic Association*, 36, 243-253 [sound samples available on-line for the *IPA* members: http://web.uvic.ca/ling/resources/ipa/members].

Fowler, C. (1977), *Timing control in speech production*, Bloomington: Indiana University Linguistic Club.

Frota, S., Vigário, M. & Martins, F. (2002), Language Discrimination and Rhythm Classes: Evidence from Portuguese, in *Proceedings of Speech Prosody 2002* (B. Bel & I. Marlien, editors), Aix-en-Provence, France, 315-318.

Frota, S., Vigário, M. & Freitas, M.J. (2003), From Signal to Grammar: Rhythm and the Acquisition of Syllable Structure, in *Proceedings of the 27th annual Boston University Conference on Language Development* (B. Beachley, A. Brown & F. Conlin, editors), Somerville, MA: Cascadilla Press, 809-821.

Fougeron, C. & Smith, C.L. (1999), French, in *Handbook of the International Phonetic Association,* Cambridge: Cambridge University Press, 78-81.

Ghazali, S., Hamdi, R. & Barkat M. (2002), Speech Rhythm Variation in Arabic Dialects, in *Proceedings of Speech Prosody 2002* (B. Bel & I. Marlien, editors), Aix-en-Provence, France, 331-334.

Gibbon, D. & Gut, U. (2001), Measuring speech rhythm, in *Proceedings of Eurospeech 2001*, Aalborg, Denmark, 95-98.

Giordano, R. (2008), On the phonetics of rhythm of Italian: patterns of duration in pre-planned and spontaneous speech, in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, May 6-9, 2008, 74-77.

Goldstein, L., Chitoran, I. & Selkirk, E. (2007), Syllable structure as coupled oscillator modes: Evidence from Georgian vs. Tashlhiyt Berber, in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, August 6-10, 2007, 241-244.

Grabe, E. (2002), Variation Adds to prosodic Typology, in *Proceedings of Speech Prosody 2002* (B. Bel & I. Marlien, editors), Aix-en-Provence, France, 127-132.

Grabe, E. & Low, E.L. (2002), Durational Variability in Speech and the Rhythm Class Hypothesis, in *Papers in Laboratory Phonology 7* (C. Gussenhoven & N. Warner, editors), Berlin: Mouton de Gruyter, 515-546.

't Hart, J., Collier, R. & Cohen, A. (1990), *A perceptual study of intonation*, Cambridge: Cambridge University Press.

Hayes, B. (1984), The Phonology of Rhythm in English, *Linguistic Inquiry*, 15, 33-74.

IPA (1999), *Handbook of the International Phonetic Association*, Cambridge: Cambridge University Press (sound samples available on-line: http://web.uvic.ca/ling/resources/ipa/handbook.htm).

Jian, H. (2004), On the syllable timing in Taiwan English, in *Proceedings of Speech Prosody 2004*, Nara, Japan, 247-250.

Keller, E. & Port, R. (2007), Speech timing: Approaches to speech rhythm, *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, August 6-10, 2007, 327-329.

Krull, D. & Engstrand, O. (2003), Speech rhythm – intention or consequence? Cross-language observations on the hyper/hypo dimension, *Phonum*, 9, 133-136, http://www.ling.su.se/fon/perilus/2003_10.pdf (last accessed 30/06/2009).

Lee, C.S. & McAngus Todd, N. (2004), Towards an auditory account of speech rhythm: application of a model of the auditory 'primal sketch' to two multi-language corpora, *Cognition*, 93, 225-254.

Lenneberg, E. (1967), *Biological foundations of language,* New York: Wiley.

Liberman M. & Prince A. (1977), On Stress and Linguistic Rhythm, *Linguistic Inquiry*, 8, 249-336 [now also in *Phonology: Critical concepts* (Ch.W. Kreidler, editor), London-New York: Routledge, 2001, 152-244].

Lindblom, Bj. & Rapp, K. (1973), Some temporal regularities of spoken Swedish, *Papers in Linguistics from the University of Stockholm*, 21, 1-59.

Lloyd James, A. (1940), *Speech signal in telephony*, London: Pitman & Sons.

MacNeilage, P.F. & Davis, B.L. (1990), Acquisition of speech production: Frames, then content, in *Attention and Performance XIII - Motor Representation and Control* (M. Jeannerod, editor), Hillsdale, NJ: LEA, 452-468.

MacNeilage, P.F. (1998), The frame/content theory of evolution of speech production, *Behavioral and Brain Sciences*, 21, 499-546.

Mairano, P. & Romano, A. (2007a), Lingue isosillabiche e isoaccentuali: misurazioni strumentali su campioni di italiano, francese, inglese e tedesco, in *Scienze vocali e del linguaggio. Metodologie di valutazione e risorse linguistiche* (V. Giordani, V. Bruseghini & P. Cosi, editors)*,* Atti del 3° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, 29 novembre – 1 dicembre 2006, Povo (Trento), Torriana (RN): EDK editore, 119-134.

Mairano, P. & Romano, A. (2007b), Inter-Subject Agreement in Rhythm Evaluation for Four Languages (English, French, German, Italian), in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, August 6-10, 2007, 1149-1152.

Mairano, P. & Romano, A. (2008a), *A comparison of four rhythm metrics for six languages,* Poster presented at the workshop 'Empirical Approaches to Speech Rhythm' (EASR), University College London, 2008.

Mairano, P. & Romano, A. (2008b), Distances rythmiques entre variétés romanes, in *La variation diatopique de l'intonation dans le domaine roumain et roman* (A. Turculeț, editor), Proceedings of the international symposium, Iași, Romania, 2008, Iași: Univ. A.I. Cuza, 251-272.

Mairano, P. & Romano, A. (2010), Un confronto tra diverse metriche ritmiche usando Correlatore, in *La dimensione temporale del parlato* (S. Schmid, M. Schwarzenbach & D. Studer, editors), Atti del 5° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Zurigo, Svizzera, 4-6 febbraio 2009 (this volume).

Marotta, G. (1985), *Modelli e misure ritmiche: la durata vocalica in italiano*, Bologna: Zanichelli.

Martinéz Celdrán, E., Fernández Planas, A.M. & Carrera Sabaté, J. (2003), Castilian Spanish, *Journal of International Phonetic Association*, 33, 255-259 (sound samples available on-line for the *IPA* members: http://web.uvic.ca/ling/resources/ipa/members).

Mehler, J., Dommergues, J., Frauenfelder, U. & Segui, J. (1981), The syllable's role in speech segmentation, *Journal of Verbal Learning and Verbal Behavior*, 20, 298-305.

Mendicino, A. & Romito, L. (1991), «Isocronia» e «base di articolazione»: uno studio su alcune varietà meridionali, *Quaderni del Dipartimento di Linguistica dell'Università della Calabria*, S. L. 3, 49-67.

Miller, M. (1984), On the perception of rhythm, *Journal of Phonetics*, 12, 75-83.

Mok, P.P.K. & Dellwo, V. (2008), Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English, in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, May 6-9, 2008, 63-66.

Molinu, L. & Romano, A. (1999), La syllabe dans un parler roman de l'Italie du Sud (variété salentine de Parabita – Lecce), in *Actes du Colloque des Journées d'Etudes Linguistiques: 'SyllabeS',* Nantes, 25-27 mars 1999*,* Nantes, France, 148-153.

Nazzi, T., Bertoncini, J. & Mehler, J. (1998), Language Discrimination by Newborns: towards an understanding of the role of rhythm, *Journal of Experimental Psychology: Human Perception and Performance*, 24, 756-766.

Nespor, M. (1993), *Fonologia*, Bologna: Il Mulino.

Nooteboom, S. (1997), The prosody of speech: melody and rhythm, in *The Handbook of Phonetic Sciences* (W.J. Hardcastle & J. Laver, editors), Oxford: Blackwell, 640-673.

O'Dell, M. & Nieminen, T. (1999), Coupled oscillators model of speech rhythm, in *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, USA, August 1-7, 1999, 1075-1078.

Pamies Bertrán, A. (1999), Prosodic Typology: On the Dichotomy between *Stress-Timed* and *Syllable-Timed* Languages, *Language Design*, 2, 103-130.

Pike, K.L. (1945), *The Intonation of American English*, Ann Arbor: University of Michigan Press.

Ramus, F. (2002), Acoustic Correlates of Linguistic Rhythm: Perspectives, in *Proceedings of the International Conference Speech Prosody 2002* (B. Bel & I. Marlien, editors), Aix-en-Provence, France, 115-120.

Ramus, F., Nespor, M. & Mehler, J. (1999), Correlates of linguistic rhythm in the speech signal, *Cognition*, 73, 265-292.

Rhardisse, N. & Abry, C. (1995), Mandible as syllable organizer, in *Proceedings of the 13th International Congress of Phonetic Sciences,* Stockholm, Sweden, 3, 556-559.

Roach, P. (1982), On the distinction between 'stress-timed' and 'syllable-timed' languages, in *Linguistic Controversies* (D. Crystal, editor), London: Arnold, 73-79.

Roach, P. (editor) (2003), *A Bibliography of Timing and Rhythm in Speech* (last updated 2nd April 2003), University of Reading, http://www.personal.rdg.ac.uk/~llsroach/timing.pdf [last accessed 30/06/2009].

Roach, P. (2004), British English: Received Pronunciation, *Journal of the International Phonetic Association*, 34, 239-245 (sound samples available on-line for the *IPA* members: http://web.uvic.ca/ling/resources/ipa/members).

Romano, A. (1999), *Analyse des structures prosodiques des dialectes et de l'italien régional parlés dans le Salento: approche linguistique et instrumentale*, Thèse de Doctorat de l'Université Stendhal de Grenoble (superv. Michel Contini) (part. published in 2001, Lille: Presses Univ. du Septentrion).

Romano, A. (2003), Accento e intonazione in un'area di transizione del Salento centro-meridionale, in *Storia politica e storia linguistica dell'Italia meridionale* (P. Radici Colace, G. Falcone & A. Zumbo, editors), *Atti del convegno internazionale di studi parlangeliani*, Messina, Italy, 2000, Messina-Napoli: Edizioni Scientifiche Italiane, 169-181.

Romano, A., Mairano, P. & Pollifrone, B. (2010), Variabilità ritmica di varietà dialettali del Piemonte in *La dimensione temporale del parlato* (S. Schmid, M. Schwarzenbach & D. Studer, editors), Atti del 5° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Zurigo, Svizzera, 4-6 febbraio 2009 (this volume).

Romito, L. & Trumper, J. (1993), Problemi teorici e sperimentali posti dall'isocronia, *Quaderni del Dipartimento di Linguistica dell'Università della Calabria*, S. L. 4, 10, 89-118.

Russo, M. & Barry, W.J. (2008a), Measuring rhythm. A quantified analysis of Southern Italian Dialects Stress Time Parameters, in *Experimental Prosody* (A. Pamies, M.C. Amorós & J.M. Pazos, editors), Actas del IV Congreso de Fonética Experimental, Granada, Spain (= *Language Design*, special issue 2), 315-322.

Russo, M. & Barry, W.J. (2008b), Isochrony reconsidered. Objectifying relations between rhythm measures and speech tempo, in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, May 6-9, 2008, 52-55.

Sachs, C. (1953), *Rhythm and tempo: a study in music history*, London: Dent.

Schmid, S. (1996), A typological view of syllable structure in some Italian dialects, in *Certamen Phonologicum III* (P.M. Bertinetto, L. Gaeta, G. Jetchev & D. Michaels, editors),

Papers from the Third Cortona Phonology Meeting, April 1996, Torino: Rosenberg & Sellier, 247-265.

Schmid, S. (2001), Un nouveau fondement phonétique pour la typologie rythmique des langues, *Poster présenté au 10ème anniversaire du laboratoire d'analyse informatique de la parole (LAIP)*, Université de Lausanne.

Schmid, S. (2004), Une approche phonétique de l'isochronie dans quelques dialectes italo-romans, in *Nouveaux départs en phonologie* (T. Meisenburg & M. Selig, editors), Tübingen: Narr, 109-124.

Schmid, S. (2008), *Measuring the rhythm of Italian dialects*, Poster presented at the workshop 'Empirical Approaches to Speech Rhythm' (EASR), University College London, 2008.

Scott, D., Isard St.D. & de Boysson-Bardies, B. (1986), On the measurement of rhythmic irregularity: a reply to Benguerel, *Journal of Phonetics*, 14, 327-330.

Sock, R., Löfqvist, A. & Perrier, P. (1996), Kinematic and acoustic correlates of quantity in Swedish and Wolof: a cross-language study, in *Proceedings of the 1st ETRW on Speech Production Modeling 'From Control Strategies to Acoustics'*, Autrans, France, 81-84.

Vayra, M., Avesani, C. & Fowler, C. (1984), Patterns of temporal compression in spoken Italian, *Proceedings of the 10th International Congress of Phonetic Sciences*, Utrecht, The Netherlands (1983), Vol. 2, 541-546.

Volín, J. (2005), Rhythmical properties of polysyllabic words in British and Czech English, in *Patterns* (J. Cermák, A. Klégr, M. Malá & P. Šaldová, editors), Praha: Kruh moderních filologu, 279-292.

White, L., Mattys, S.L., Series, L. & Gage, S. (2007), Rhythm Metrics Predict Rhythmic Discrimination, in *Proceedings of the 16th International Congress of Phonetic Sciences,* Saarbrücken, Germany, August 6-10, 2007, 1009-1012.

11. APPENDIX. SOURCES OF THE SPEECH SAMPLES ANALYSED IN 6.1

Original data for several languages have been recorded at the *LFSAG* within the framework of various research projects or by individual dissertations on different topics (in parentheses we summarise the reference to the *project name* or to the author of the recording).

French_european	*IPA* (1999)
Romanian_muntenian	*LFSAG* (H. Bandea, 2006)
Italian_1	*LFSAG* (A. Romano - P. Mairano, 2006)
Finnish_2	*LFSAG* (L. Capovilla, 2007)
Chinese_Mandarin	*LFSAG* (S. Pittoni, 2008)
Icelandic_10	*LFSAG* (P. Mairano, 2006)
Icelandic_4	*LFSAG* (P. Mairano, 2006)
Italian_2	Canepari (2004-)
Finnish_1	*LFSAG* (L. Capovilla, 2007)
Romanian_moldavian	*LFSAG* (A. Romano, 2007)
French_canadian	*LFSAG* (*Loquendo*, 2006)
Italian_3	Costamagna (2000)
Portuguese_brazilian (SP)	Barbosa Almeida & Albano Cavalcanti (2004)
Italian_4	*LFSAG* (L. Calabrò, 2009)
Spanish_castilian	Martinéz Celdrán, Fernández Planas & Carrera Sabaté (2003)
Japanese_fast	*LFSAG* (A. Romano - P. Mairano, 2006)
English_GA	*IPA* (1999)
Japanese_slow	*LFSAG* (A. Romano – P. Mairano, 2006)
Chinese_Cantonese	*LFSAG* (S. Pittoni, 2008)
Turkish	*LFSAG* (A. Romano – P. Mairano, 2006)
Russian	*LFSAG* (*CELI*, 2008)
Czech	*LFSAG* (D. Brdičko, 2007)
English_RP	Roach (2004)
English_Aus	*LFSAG* (*CELI*, 2009)
Portuguese_european	*IPA* (1999)
German_IPA	*IPA* (1999)
Zurich_German	Fleischer & Schmid (2006)
German_1	*LFSAG* (L. Capovilla, 2007)
German_2	*LFSAG* (L. Capovilla, 2007)
Arabic_lebanese	*LFSAG* (*CELI*, 2009)

LINGUISTICA, FONETICA E FONOLOGIA

UN CONFRONTO TRA DIVERSE METRICHE RITMICHE USANDO CORRELATORE

Paolo Mairano, Antonio Romano
Laboratorio di Fonetica Sperimentale 'Arturo Genre', Università degli Studi di Torino
paolo.mairano@unito.it, antonio.romano@unito.it

1. SOMMARIO

Basandosi sulle teorie di percezione del parlato da parte dei bambini (si veda Mehler *et al.*, 1996), Ramus *et al.* (1999) hanno proposto tre correlati ritmici (ΔC, ΔV and %V) che permetterebbero di distinguere i due gruppi di lingue. Questo nuovo approccio alla tipologia ritmica ha dato un impulso alla ricerca in questo campo cosicché, come è noto, sono stati proposti nuovi correlati: i PVI di Grabe & Low (2002), i Varco di Dellwo & Wagner (2003 e seguenti) e, recentemente, i CCI di Bertinetto & Bertini (2008).

Un primo obiettivo che ci siamo posti in questo lavoro è stato quello di testare il mutamento dei risultati al variare di alcuni fattori: sono stati calcolati %V, ΔV, ΔC, nPVI(V), rPVI(C), Varco(V), Varco(C), CCI(V) e CCI(C) per 36 campioni de *Il vento di tramontana e il sole* in varie lingue. Questa scelta rispecchia la nostra convinzione che sia opportuno studiare campioni di parlato controllato prima di affrontare il parlato spontaneo e che non sia necessario utilizzare campioni estremamente lunghi poiché alcuni test di discriminazione hanno dimostrato che il cervello umano distingue lingue isosillabiche e isoaccentuali anche con campioni di pochi secondi (v. Ramus *et al.*, 1999). Proponiamo anche di evitare il ricorso alla segmentazione automatica in quanto il guadagno in termini di tempo è controbilanciato da una perdita di precisione; al contrario, abbiamo pensato di automatizzare il processo di calcolo dei correlati: a questo scopo l'autore PM ha realizzato *Correlatore*, disponibile sul sito del Laboratorio di Fonetica Sperimentale 'Arturo Genre' di Torino; si tratta di un programma in Tcl/Tk che calcola i valori di %V, ΔV, ΔC, nPVI(V), rPVI(C), Varco(V), Varco(C), CCI(V) e CCI(C) e costruisce i grafici a partire dai file di annotazione di *Praat* (TextGrid) etichettati semplicemente come CV (consonante-vocale) o in SAMPA.

I dati ottenuti ci permettono di analizzare il variare dei risultati a seconda: a) dei diversi correlati utilizzati; b) dei diversi parlanti di una stessa lingua; c) di chi segmenta. Non sono stati ancora debitamente valutati gli eventuali effetti della velocità d'eloquio (argomento discusso in numerosi studi, v. per es. Dellwo & Wagner, 2003) perché, comunque, i nostri campioni sono piuttosto omogenei a questo riguardo (5-6,5 syll/s). In tutti i casi, si nota un certo grado di sovrapposizione tra gruppi di lingue che graviterebbero tradizionalmente attorno ai due poli sillabico e accentuale, presumibilmente dovuto al fatto che i diversi correlati sono sensibili a fenomeni differenti. Tuttavia, è difficile stabilire quali correlati rispecchino meglio la tradizionale dicotomia di lingue isosillabiche/isoaccentulali in quanto non sembra possibile avere un riscontro oggettivo del punto esatto in cui un determinato campione debba situarsi all'interno del continuum.

2. INTRODUZIONE

Negli ultimi anni, in seguito alla pubblicazione di Ramus *et al.* (1999), che proponeva l'utilizzo di tre correlati acustici (%V, ΔV, ΔC) al fine di differenziare i tradizionali gruppi ritmici detti isosillabico e isoaccentuale, si è assistito a un proliferare di studi in questo

campo, che hanno portato anche alla proposta di nuovi correlati del ritmo: in particolare, i PVI di Grabe & Low (2002), i Varco di Dellwo & Wagner (2003 e seguenti) e, recentemente, i CCI di Bertinetto & Bertini (2008).

Nel corso di questo lavoro, dopo un breve sunto dei lavori in questo campo, presenteremo *Correlatore*, uno strumento sviluppato dall'autore PM volto ad automatizzare il calcolo dei correlati e reso disponibile sul sito del Laboratorio di Fonetica Sperimentale 'Arturo Genre' di Torino. In seguito, verranno presentati i risultati delle metriche ritmiche che abbiamo ottenuto su dati di parlato letto di un numero cospicuo di lingue (comunque decisamente superiore ad altri studi finora pubblicati in questo campo).

3. LINGUE ISOSILLABICHE E LINGUE ISOACCENTUALI

Anche se, come notato da diversi autori (cfr. per es. Eriksson, 1991), la differenza tra diversi tipi ritmici di lingue era già stata rimarcata precedentemente, la dicotomia *stress-timed languages* vs. *syllable-timed languages* si fa tradizionalmente risalire a Pike (1945).[1] Abercrombie (1967) introdusse il concetto di isocronia, ipotizzando l'esistenza di due tipologie ritmiche (isoaccentuale e isosillabica) a cui necessariamente apparterrebbero tutte le lingue del mondo. Le lingue isoaccentuali (come l'inglese, il tedesco e alcune lingue slave) presenterebbero isocronia a livello accentuale, cioè la distanza tra un accento e l'altro sarebbe relativamente costante, provocando un adeguamento della durata delle sillabe in modo da stare entro i confini dati da due accenti; viceversa, le lingue isosillabiche (come l'italiano, il francese e lo spagnolo) presenterebbero isocronia a livello sillabico, cioè la durata delle sillabe in queste lingue tenderebbe a rimanere relativamente costante. L'appartenenza di una lingua a una determinata classe ritmica avrebbe peraltro alcune corrispondenze nel sistema metrico adottato nella letteratura tradizionale, come è dimostrato, per esempio, dai sistemi metrici in uso in inglese e italiano, il primo basato sui piedi, il secondo sulle sillabe.

Negli anni successivi (e, in realtà, anche prima di Pike, 1945, si veda per esempio Classe, 1939, citato in Bertinetto, 1989 e Eriksson, 1991), molti autori hanno cercato delle prove sperimentali a conferma di questa ipotesi, ma generalmente essa è stata smentita; fra questi, citiamo Roach (1982), il quale conclude che l'isoaccentualità e l'isosillabicità siano da attribuire esclusivamente alla percezione: "a language is syllable-timed if it *sounds* syllable-timed" (Roach, 1982:78). Per ulteriori dettagli riguardo a questi studi rimandiamo a Bertinetto (1989), che, inoltre, ribadisce quanto già affermato in precedenza (Bertinetto, 1981) e da Dauer (1983) riguardo all'esigenza di considerare l'isoaccentualità e l'isosillabicità non come due categorie in rapporto reciprocamente esclusivo, bensì come i due poli di un continuum lungo il quale si distribuiscono le lingue; inoltre, viene ipotizzato che alcune caratteristiche fonologiche giochino un ruolo primario nella classificazione ritmica delle lingue. In particolare, vengono citate a) la presenza di un accento mobile vs. fisso; b) la presenza vs. l'assenza di fenomeni macroscopici di riduzione vocalica; c) una complessa

[1] Eriksson fa notare come il fonetista del XVIII sec. Joshua Steele avesse già ipotizzato che in inglese gli accenti occorressero a intervalli regolari; naturalmente, la sua teoria era basata esclusivamente su criteri impressionistici, data la mancanza all'epoca di strumentazione adeguata a verificare una simile ipotesi. Più avanti, Lloyd James (1940, citato dallo stesso Pike, 1945), propose una distinzione tra lingue con un ritmo a mitragliatrice (*machine-gun rhythm*) e lingue a codice Morse (*Morse code rhythm*).

vs. semplice struttura sillabica; d) la tendenza più o meno accentuata delle sillabe di attrarre nuovo 'materiale fonologico' per formare sillabe più complesse e/o pesanti.[2] Queste ipotesi hanno rinvigorito gli studi in questo campo e in Italia numerose successive indagini hanno approfondito la struttura ritmica e l'organizzazione temporale dell'italiano, delle sue varietà e di alcuni dialetti d'Italia (si vedano Bertinetto, 1977; Vayra *et al.*, 1984; Marotta, 1985; Farnetani & Kori, 1990; Romito & Trumper, 1993).

In anni più recenti, le ipotesi avanzate da Bertinetto (1981 e 1989) e Dauer (1983) hanno spinto alcuni autori a cercare dei correlati acustici delle proprietà fonologiche citate sopra e considerate responsabili dell'organizzazione ritmica e temporale di una lingua. Ramus, Nespor & Mehler (1999) (d'ora in poi Ramus *et al.*, 1999) hanno proposto l'utilizzo di ΔV (deviazione standard degli intervalli vocalici, indicativa della presenza/assenza di riduzione vocalica) e ΔC (deviazione standard degli intervalli intervocalici, indicativa di una struttura sillabica complessa) e %V (percentuale vocalica, indicativa di entrambe le proprietà); applicando queste formule a brevi campioni di otto lingue, i tre autori hanno ottenuto risultati confortanti, con valori più bassi di ΔV e ΔC per lingue come l'italiano, lo spagnolo e il francese, più alti per lingue come l'inglese e l'olandese. In studi immediatamente successivi, è emersa l'estrema sensibilità di questi tre parametri rispetto alla velocità d'eloquio e alla sua variazione; al fine di mitigare gli effetti di questa variabile, Dellwo & Wagner (2003) e poi Dellwo (2006) hanno sviluppato i VarcoV e VarcoC, che prevedono la divisione del risultato della deviazione standard per la durata media degli intervalli. Parallelamente a Ramus *et al.* (1999) e sulla base di presupposti simili, Grabe & Low (2002) hanno invece proposto l'utilizzo dei *PVI* (*Pairwise Variability Index*) sui valori di durata degli intervalli consonantici e vocalici (a questi ultimi viene applicata una formula leggermente diversa al fine di normalizzare rispetto agli effetti della velocità d'eloquio, gli *nPVI*) e anch'essi hanno ottenuto dei risultati conformi alle aspettative; il vantaggio di questo approccio è che, a differenza della formula della deviazione standard, esso tiene conto della successione temporale dei segmenti.

Nonostante le numerose critiche a questo approccio (v. tra gli altri Barry & Russo, 2003), i correlati hanno goduto di grande successo e sono stati recentemente utilizzati in molti studi al fine di ottenere una valutazione ritmica di campioni sonori. Ancora più recentemente, Bertinetto & Bertini (2008) hanno proposto un nuovo indice per la tipologizzazione ritmica, i CCI (*Control and Compensation Index*), di natura parzialmente diversa rispetto ai suoi fratelli. Oltre a essere un modello "phonologically driven" (secondo le parole degli autori), esso si ricollega a studi precedenti sulla compensazione (molti dei quali risalenti a Fowler, 1977). La formula dei CCI riprende quella dei PVI, ma viene

[2] Schmid (1996) ha poi aggiunto alla lista altre caratteristiche fonologiche che possono avere un'influenza: l'inventario dei tipi sillabici (generalmente caratterizzato da pochi tipi per le lingue isosillabiche); la generale preferenza delle lingue isosillabiche per i tipi CV; una distribuzione e una casistica di occorrenza grosso modo invariabili dei tipi sillabici in posizione forte o debole nelle lingue isosillabiche; la presenza nelle lingue isoaccentuali di un maggior numero di gruppi consonantici complessi e spesso di sonoranti sillabiche; la mancanza nelle lingue isoaccentuali di opposizioni scempia-geminata; una tendenza più accentuata nelle lingue isoaccentuali all'allungamento delle vocali accentate e al frangimento dei nuclei, insieme a una maggiore predisposizione alla centralizzazione delle vocali atone (e alla neutralizzazione tra i timbri in posizione non accentata).

applicata agli intervalli vocalici e consonantici divisi per il numero di segmenti fonologici di cui sono composti con lo scopo di misurare li livello di compressione permesso da una lingua;[3] il supposto di base è infatti che le lingue tradizionalmente considerate isoaccentuali permettano un maggior livello di compressione rispetto alle lingue considerate isosillabiche, le quali mantengono invece un maggior 'controllo' sulla durata dei segmenti. I due autori hanno condotto uno studio sui materiali dell'archivio *API*, ottenendo risultati che hanno corroborato le ipotesi. Questo è invece il primo studio in cui le formule dei CCI vengono applicate a campioni non italiani.

Prima di continuare, ci sembra conveniente aprire una breve parentesi terminologica riguardo ai termini da adottare per le due tipologie ritmiche di lingue, dato che la etichette utilizzate nella letteratura non sono uniformi e, soprattutto, rischiano di essere ambigue. Se da un lato le originarie (nonché originali) espressioni 'lingue a mitragliatrice' e 'lingue a codice Morse' suonano decisamente colorite e un po' *naïf*, i termini *stress-timed* e *syllable-timed*, e in misura ancora maggiore, 'isoaccentuale' e 'isosillabico' sono troppo legati all'osteggiata ipotesi di isocronia, ormai da tempo smentita. Come denominare, dunque, i due gruppi di lingue? Se le ipotesi di Bertinetto & Bertini (2008) continueranno a dare buoni frutti, e i dati di lingue 'isosillabiche' confermeranno effettivamente che si tratta di lingue 'a controllo' mentre i dati di lingue 'isoaccentuali' confermeranno che si tratta di lingue 'a compensazione', queste nuove categorie potrebbero sovrapporsi a quelle tradizionali, portando anche a una sostituzione terminologica. Per il momento, preferiamo non spingerci così avanti, mantenendo una differenziazione tra le categorie tradizionali e il nuovo approccio. Questo è utile anche per enfatizzare la diversa natura dei fenomeni rispecchiati da un lato dai delta, dai varco e dai PVI, dall'altro dai CCI. Quindi, utilizzeremo i termini 'lingue a compensazione' e 'lingue a controllo' riferendoci ai CCI, e 'lingue sillabiche' e 'lingue accentuali' riferendoci agli altri correlati, evitando quindi il prefissoide *iso-*, colpevole di richiamare le smentite ipotesi di isocronia.

4. CORRELATORE

Gli ormai numerosi studi che utilizzano questi correlati sono stati in molti casi criticati per l'esiguità dei dati utilizzati, spesso composti da qualche frase pronunciata da un numero limitato di parlanti (si vedano gli stessi Ramus *et al.*, 1999, nonché, seppur in misura minore, Grabe & Low, 2002). Questo è certamente dovuto al fatto che il calcolo dei correlati è estremamente dispendioso in termini di tempo, che richiede un'etichettatura di tutti i dati (con tutti i problemi connessi, data anche la necessità di un'interpretazione su base fonetica o fonologica di ogni segmento come V o C – si veda Mairano & Romano, 2007b, per maggiori dettagli su questo punto) e il trasferimento delle misure su dei fogli di calcolo. Inoltre, come si può facilmente intuire, con l'aumentare dei campioni e con l'utilizzo di nuovi correlati, aumentano anche i problemi di organizzazione dei dati, particolarmente notevoli per chi voglia lavorare sui CCI. Questi, infatti, richiedono un'etichettatura leggermente diversa rispetto a quella necessaria per gli altri correlati (per esempio, è necessario separare gli iati nonché tenere conto del numero di segmenti di cui è composto ogni intervallo); di conseguenza, chi disponesse di dati etichettati secondo i criteri esposti da

[3] Gli autori forniscono una discussione dettagliata sui casi problematici in merito a ciò che deve essere considerato come segmento singolo o multiplo (consonanti e vocali geminate, iati, …). Si rimanda quindi a Bertinetto & Bertini (2008).

Ramus *et al.* (1999) o Grabe & Low (2002), si troverebbe a dover rivedere le etichettature e creare nuovi fogli di calcolo con i nuovi dati. Quindi, dopo aver sperimentato in un primo momento queste problematiche, abbiamo deciso di automatizzare il calcolo dei correlati, cioè di sviluppare uno script che calcolasse tutti i correlati a partire da file che avevamo etichettato con *Praat*. In seguito, avendo notato il notevole vantaggio di questo strumento, abbiamo pensato di sviluppare un vero e proprio programma che permettesse non solo il calcolo dei correlati, ma anche un'organizzazione dei risultati e una creazione veloce di grafici. Desideriamo anche far presente che, come già spiegato in pubblicazioni precedenti (si vedano, ad esempio, Mairano & Romano, 2007a e b), abbiamo deciso di optare ancora per la segmentazione manuale al fine di evitare che i risultati siano condizionati dagli errori commessi dagli strumenti di segmentazione automatica, il che porterebbe a una perdita di precisione potenzialmente notevole.[4] Tuttavia, teniamo comunque a precisare che l'automatizzazione del calcolo dei correlati non è in opposizione all'automatizzazione della segmentazione: i due approcci potrebbero essere combinati da chi voglia acquisire il maggior numero di dati con il minor dispendio di tempo possibile (a spese però della precisione).

Correlatore è un programma sviluppato in Tcl/Tk dall'autore PM al fine di automatizzare il calcolo dei correlati ritmici più utilizzati negli ultimi anni, in particolare %V, ΔV e ΔC (Ramus *et al.*, 1999), i varco (Dellwo & Wagner, 2003, e Dellwo, 2007), i PVI (rPVI e nPVI – v. Grabe & Low, 2002) e i CCI (Bertinetto & Bertini, 2008).[5] Esso è distribuito in licenza GPL ed è disponibile sul sito del Laboratorio di Fonetica Sperimentale 'Arturo Genre' di Torino, all'indirizzo web http://www.lfsag.unito.it/correlatore/.

Di seguito ne verranno esplicati il funzionamento, le caratteristiche e i vantaggi; la descrizione è aggiornata alla versione 2.0, mentre alcune differenze rispetto alla versione 1.0 sono riportate in nota. Una documentazione dettagliata è comunque disponibile alla pagina web http://www.lfsag.unito.it/correlatore/readme.html.

---

[4] Va notato che la segmentazione manuale non è comunque di per sé garanzia di precisione, tanto è vero che, come dimostrato da Mairano & Romano (2007a e b), etichettatura fatte da annotatori diversi possono portare a risultati diversi (anche a causa di differenti scelte di classificazione fonologica).

[5] Riguardo alla scelta di Tcl/Tk, essa potrebbe stupire a causa del fatto che questo linguaggio di programmazione viene ormai considerato come vecchio e desueto. Tuttavia, possiede alcuni vantaggi che ce l'hanno fatto preferire ad altri: in particolare, 1) è *open-source*; 2) è multipiattaforma, cosa che consente ai sorgenti di *Correlatore* di funzionare senza modifiche su qualsiasi piattaforma supportata da Tcl/Tk 8.5; 3) grazie a Tclkit è possibile creare degli eseguibili leggeri in termini di risorse richieste al sistema che permettono il funzionamento di *Correlatore* anche laddove non sia presente un'installazione di Tcl/Tk; 4) Tcl mette a disposizione le espressioni regolari (molto utili per l'analisi dei TextGrid, che sono file di testo), Tk fornisce un *canvas* (un widget che ha permesso di sviluppare abbastanza facilmente un modulo per la creazione di grafici), mentre l'estensione Img permette di convertire i grafici in diversi formati immagine; 5) l'utilizzo di Tcl/Tk permette un'eventuale futura integrazione di *Correlatore* con lo *Snack Sound Toolkit* (di Kåre Sjölander). Inoltre, riportiamo che Tk è stato spesso criticato a causa del fatto che le interfacce grafiche create con questo toolkit non sono esteticamente apprezzabili e non si intergrano nei diversi sistemi operativi; se anche ciò sia mai stato di qualche rilevanza, non lo è più a partire dalla versione 8.5, che utilizza i controlli nativi su *Windows* e Mac OS X (non ancora su Unix – ma l'aspetto è stato migliorato anche qui).

Correlatore calcola i correlati a partire dai file di annotazione prodotti con *Praat* (ovvero i TextGrid). Quindi, è sufficiente etichettare un file sonoro con *Praat* per ottenere le misure di tutti i correlati citati sopra e, eventualmente, costruire grafici con i risultati, senza bisogno di creare enormi fogli di calcolo.

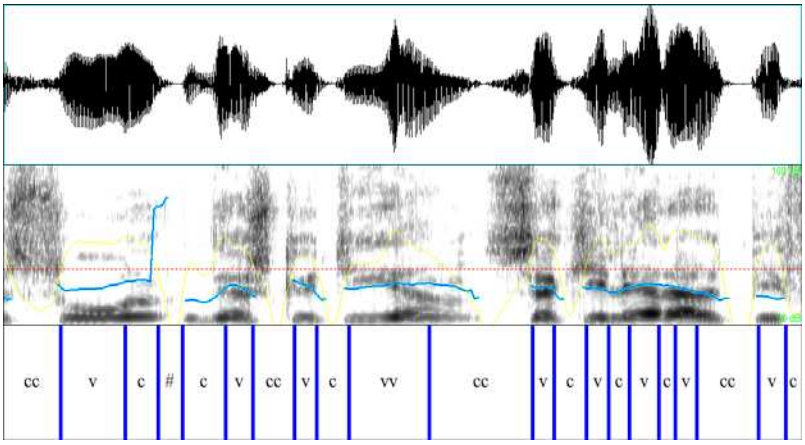


Figura 1: Esempio di annotazione CV

Per la sua utilizzazione, il primo passo consiste nell'etichettare un file sonoro con *Praat*: l'etichettatura può essere fatta come CV (consonante-vocale) o in SAMPA, ma in entrambi i casi è necessario seguire alcune annotazioni affinché tutti i correlati vengano calcolati correttamente. Nel caso di un'etichettatura CV è necessario creare un'etichetta per ogni intervallo vocalico o consonantico e annotarla con tante 'c' o 'v' quanti sono i segmenti fonologici che compongono l'intervallo. Per esempio, 'marcio' deve essere etichettato come |c|v|cc|v|, 'palla' come |c|v|cc|v|, 'accipicchia' come |v|cc|v|c|v|ccc|v|. Le pause vanno lasciate vuote o etichettate come #. Questo tipo di trascrizione lascia l'utente libero di decidere se considerare vocalici o consonantici i segmenti dubbi (per esempio le consonanti sillabiche e i *glide*)[6] e di controllare pienamente la suddivisione degli intervalli; in questo modo si possono seguire le istruzioni di Bertinetto & Bertini (2008) per il calcolo dei CCI: ad esempio, gli iati possono essere etichettati come 2 intervalli distinti: 'suo' |c|v|v|. In figura 1 è mostrato un esempio.[7]

[6] Per quanto alcuni suoni approssimanti possano presentare caratteristiche vocaliche in alcune varietà (soprattutto in prossimità della vocale seguente, nel caso dei dittonghi ascendenti), si tratta di segmenti generalmente considerati appartenenti all'attacco consonantico (e così sono stati da noi trattati). L'utente di *Correlatore* è comunqe libero di optare per una diversa interpretazione, modificando la variabile di sostituzione (v. *infra*).

[7] In alternativa è anche possibile utilizzare un'etichettatura più semplice in cui non venga indicato il numero di segmenti che compone ogni intervallo, es. 'palla' |c|v|c|v|, 'accipicchia' |v|c|v|c|v|c|v|, ma questo porterà a risultati erronei dei CCI (la cui formula prevede una divisione per il numero di segmenti di ogni intervallo, che in questo caso risulterebbe sempre 1, dando lo stesso risultato degli rPVI), mentre i risultati degli altri correlati rimarranno invariati.

In alternativa, è possibile optare per un'etichettatura SAMPA. *Correlatore* trasforma le trascrizioni SAMPA in una sequenza CV tramite la 'variabile di sostituzione': si tratta di una variabile che contiene tutti i simboli che devono essere considerati vocalici: cioè, quando il programma apre un file TextGrid etichettato in SAMPA, ogni etichetta non vuota viene sostituita con una V se contiene uno dei simboli presenti nella variabile di sostituzione, altrimenti con una C (a meno che contenga # - in qual caso viene considerata una pausa). Il valore di default della variabile è aeiouyAEIOUY@MQV&1236789={} (quindi sono incluse le consonanti sillabiche (=), mentre i *glide* vengono considerati consonantici), ma è possibile modificarlo cliccando sulla barra in basso, in cui è indicato il suo valore (v. figura 2).

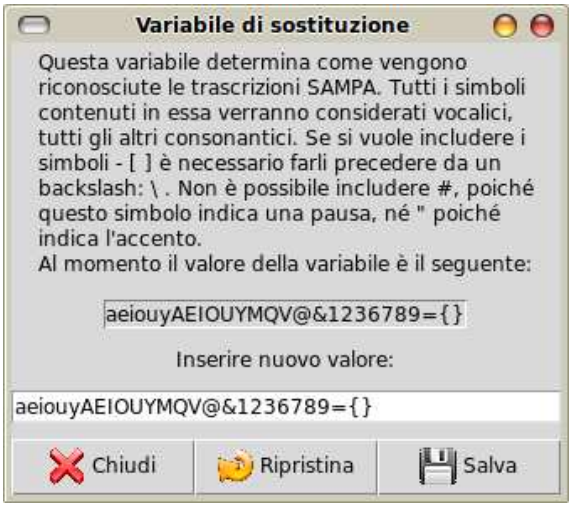


Figura 2: Finestra di Correlatore per modificare il valore della variabile di sostituzione.

Le convenzioni da seguire nel caso di un'etichettatura SAMPA sono le seguenti:

- è necessario che a ogni etichetta corrisponda un solo fono (quindi solo una vocale o una consonante, non un intervallo vocalico o consonantico);
- i fonemi fonologicamente geminati (es. le vocali lunghe del finlandese e le consonanti geminate dell'italiano) vanno annotati con due etichette distinte (nonostante il confine tra i 2 foni sia naturalmente fittizio). Ad esempio, il finlandese 'saami' deve essere etichettato |s|a|a|m|i|, e non |s|a:|m|i| né |s|aa|m|i|;
- è possibile usare i diacritici SAMPA standard, ma se si utilizzano diacritici non standard, questi potrebbero interferire con la variabile di sostituzione (v. punto seguente). Per esempio, se si utilizza t_u (invece di t_w) per indicare una occlusiva dentale sorda labializzata, quell'etichetta verrà erroneamente considerata vocalica a causa del simbolo u;
- le pause devono essere etichettate con # oppure lasciate vuote;

- durante il processo di segmentazione e calcolo dei correlati di un TextGrid etichettato come SAMPA, *Correlatore* costruirà gli intervalli vocalici e consonantici sommando le durate di ogni consonante/vocale esclusivamente secondo un criterio di adiacenza. Questo significa che gli iati (essendo 2 vocali in sequenza) verranno considerati come un solo intervallo vocalico; questo non avrà alcun effetto sui valori di percentuale vocalica, delta, varco, rPVI e nPVI, ma avrà conseguenze (presumibilmente leggere, almeno per lingue come l'italiano) sul calcolo dei CCI. Per chi desideri ottenere risultati più precisi per i CCI, si consiglia dunque una segmentazione di tipo CV (v. *supra*).[8]

Figura 3: Finestra principale di *Correlatore* 2.0

Una volta che almeno un file sonoro è stato etichettato e salvato come TextGrid, è possibile aprire *Correlatore*, di cui la figura 3 mostra la finestra principale: nel riquadro vengono mostrati i TextGrid pronti, nel frame a destra vi sono i comandi che permettono di accedere ai *report*, in basso si vede il valore corrente della variabile di sostituzione. Se l'eseguibile di *Correlatore* si trova nella stessa cartella del TextGrid, quest'ultimo verrà trovato automaticamente, altrimenti sarà necessario specificarne il percorso. A questo punto, selezionando un TextGrid e premendo sul tasto 'Segmenta e calcola correlati', si aprirà un'altra finestra (v. figura 4) in cui, dopo aver scelto il *tier* che contiene le annotazione e dopo aver specificato il tipo di notazione (SAMPA o CV), verranno riempite le caselle a lato con le durate dei singoli segmenti vocalici e consonantici e con i valori dei correlati.

[8] Uno iato in etichettatura CV sarebbe etichettato |V1|V2| e *Correlatore* mantiene separati i due intervalli vocalici per il calcolo dei CCI (es.: D_{V1} = 40 ms; D_{V2} = 80 ms); lo stesso iato in etichettatura SAMPA (ad es. con V1=[a] e V2=[e]) andrebbe etichettato |a|e| ma per *Correlatore* questo risulterebbe un unico intervallo vocalico (D=120 ms) e non sarebbe differenziato da un dittongo.

La figura 4 mostra la finestra di segmentazione di *Correlatore*. È necessario scegliere quale *tier* contiene l'annotazione e specificare se si tratta di etichettatura SAMPA o CV. I riquadri bianchi verranno riempiti dopo la segmentazione. Inoltre, nel riquadro in basso, verrà data una rappresentazione grafica degli intervalli (v. figura 5).

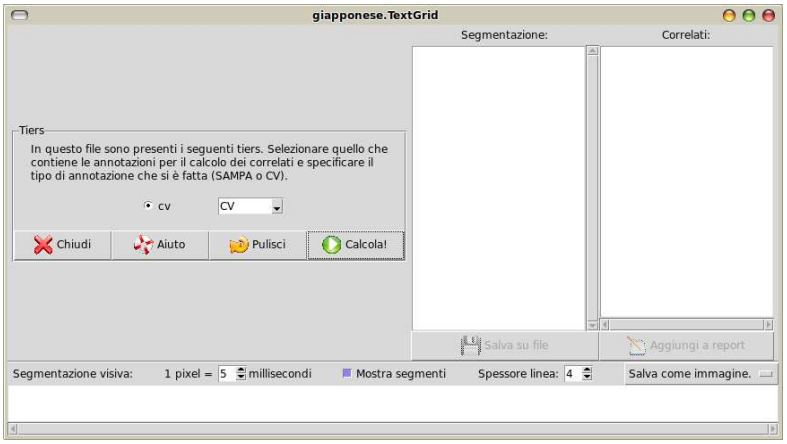


Figura 4: Finestra di segmentazione di *Correlatore*

Va notato che per ogni correlato vengono restituiti due valori: questo perché le formule vengono applicate in due diversi modi: (a) su tutti i valori di durata degli intervalli (separatamente per gli intervalli vocalici e quelli consonantici, come è ovvio) oppure (b) sui valori di durata degli intervalli di ogni segmento interpausale, poi calcolando la media dei valori così ottenuti (sempre separatamente per gli intervalli vocalici e quelli consonantici). Questo secondo metodo sembra più che altro utile per ΔC e ΔV, i quali non prevedono nessuna normalizzazione dal punto di vista della velocità d'eloquio (in effetti, gli stessi Ramus *et al.* (1999) hanno applicato le formule separatamente per ogni frase, poi calcolando la media di questi risultati parziali).

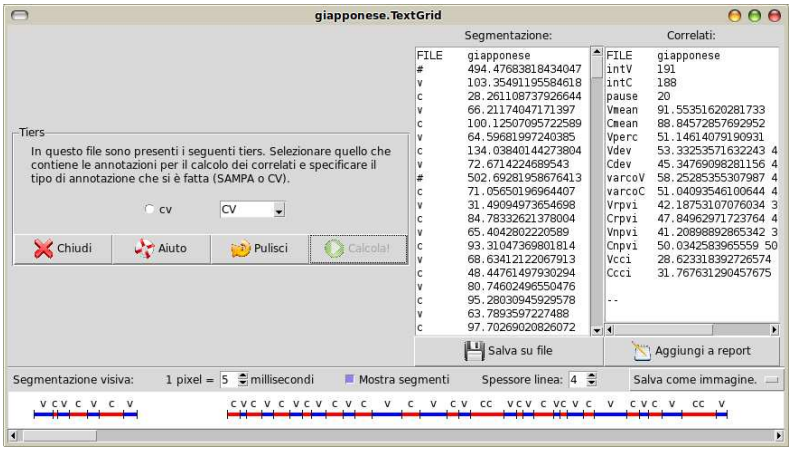


Figura 5: Finestra di segmentazione di *Correlatore* dopo i calcoli

Nella figura 5, il riquadro a sinistra mostra le durate degli intervalli vocalici/consonantici, mentre il grafico sotto né da una rappresentazione grafica. Il riquadro a destra mostra invece i valori dei correlati, che è possibile salvare premendo su 'Aggiungi a report'.

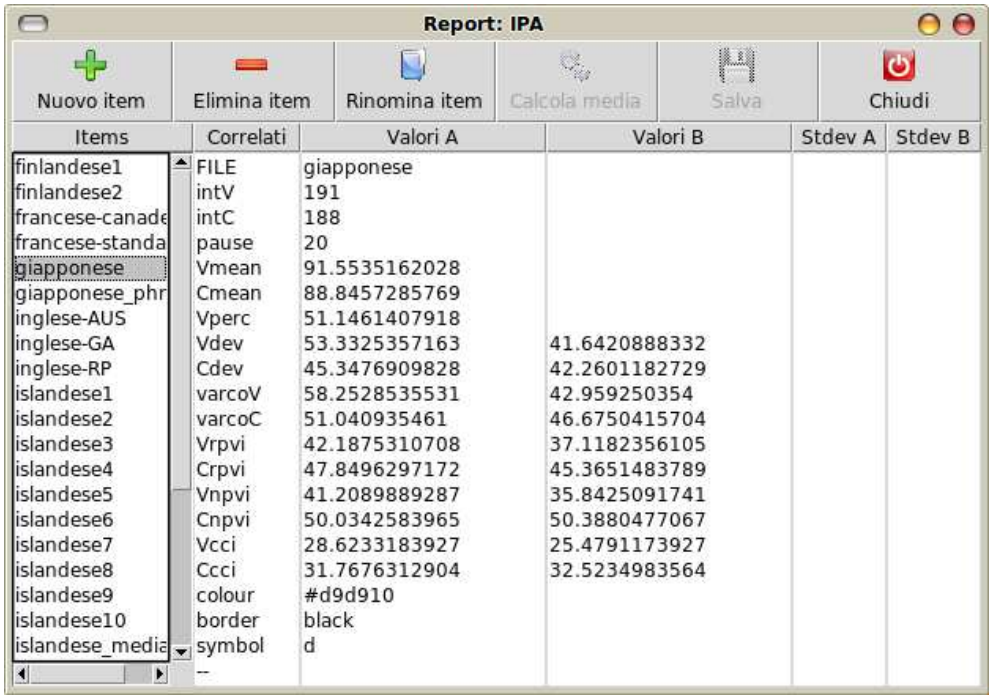


Figura 6: Finestra di visualizzazione del *report*

Una volta ottenuti i risultati dei correlati, è possibile salvarli in un *report*: il *report* è un file di testo contenuto nella cartella di configurazione di *Correlatore* che contiene i valori dei correlati e alcuni altri dati organizzati in maniera coerente.[9] Salvando i dati in un *report*, è possibile visualizzarli successivamente (v. figura 6) o utilizzarli per costruire dei grafici (v. figura 7).[10]

[9] Si vedano le istruzioni sul sito del Laboratorio di Fonetica Sperimentale 'Arturo Genre' per ulteriori dettagli a proposito.

[10] Nella versione 1.0 di *Correlatore* esisteva un solo *report*, che conteneva tutti i valori salvati. Tuttavia, ci siamo presto resi conto che, con l'aumentare dei dati, questa organizzazione diventava scomoda e rendeva difficile la creazione di grafici che contenessero solo i valori di determinati TextGrid. Per questo motivo, a partire dalla versione 2.0 di *Correlatore* è possibile creare più di un *report* e, al momento del salvataggio, è necessario specificare in quale *report* si desidera inserire i nuovi dati. Inoltre, è adesso possibile importare ed esportare i *report*, senza più bisogno di metter mano nella cartella di configurazione di *Correlatore*.

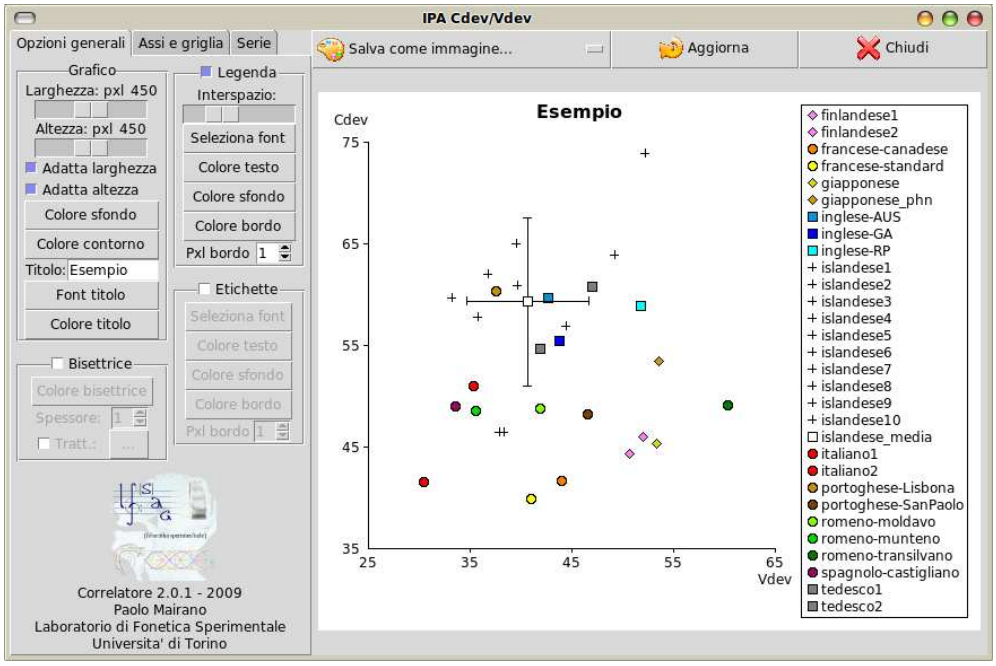


Figura 7: Finestra di costruzione dei grafici, completamente personalizzabili tramite i comandi a sinistra

La visualizzazione del *report*, oltre a consentire alcune operazioni come l'aggiunta, l'eliminazione e la rinominazione dei dati salvati, permette di calcolare la media e la deviazione standard di due o più item: questo risulta utile nel caso si abbiano più campioni di una stessa lingua (come l'islandese nel nostro caso, v. *infra*) o un campione etichettato da più persone. Verrà calcolata anche la deviazione standard, che sarà (facoltativamente) visualizzata nei grafici sotto forma di barre d'errore.

5. I NOSTRI DATI

Al fine di indagare la tipologia ritmica delle lingue che abbiamo preso in considerazione, abbiamo preferito utilizzare forme di parlato più controllate, esenti dalle discontinuità e interferenze di forme di parlato più libere. Siamo consci dei rischi che comporta questa scelta e, più in generale, l'uso di materiali prodotti in laboratorio; tuttavia, essendo l'indagine in questo campo appena agli inizi e dovendo ancora essere confermata la validità delle metriche ritmiche, riteniamo che sia indispensabile concentrarsi su un parlato più controllato prima di affrontare il parlato spontaneo e i complessi fenomeni connessi che ne rendono complicata la descrizione, data la loro (almeno potenzialmente) forte influenza sul ritmo e sulle misure a esso associate. Negli ultimi anni, in effetti, vari studi hanno calcolato le metriche ritmiche su parlato spontaneo o semi-spontaneo, raggiungendo spesso risultati non conformi alle aspettative, si veda ad esempio Barry & Russo (2003). Ci è sembrato quindi auspicabile utilizzare, come primo passo, un parlato letto di buona qualità – nel quale sono ben presenti proprietà ritmiche immediatamente percepibili – giungendo a una descrizione adeguata per questo tipo di parlato per meglio comprendere le variabili che

possono agire, interferire o limitare la validità delle metriche prese in analisi.

Solo in un secondo momento sarà adeguato prendere in considerazione altre forme di parlato e analizzarle alla luce di quanto scoperto circa le variabili in gioco. Inoltre, per quanto alcuni autori auspichino il ricorso a metodi basati su robuste tecniche di estrazione automatica dei parametri che necessitano di una cospicua quantità di dati (v. discussione in Mairano & Romano, 2007a), riteniamo che – avvicinandosi piuttosto alla definizione di opportune metriche che consentano di farlo – non sia necessario utilizzare campioni estremamente lunghi. Alcuni test di discriminazione hanno dimostrato che un ascoltatore può distinguere lingue sillabiche e accentuali anche con campioni di pochi secondi (v. Ramus *et al.*, 1999). Individuare metriche in grado di emulare queste prestazioni costituisce un obiettivo ottimale.

Finora abbiamo utilizzato 36 registrazioni della storia *La tramontana e il sole*, utilizzata anche dall'*International Phonetic Association* per le illustrazioni del *Handbook of the IPA*. Di questi, 7 (inglese britannico, inglese americano, inglese australiano, francese standard, tedesco standard, portoghese europeo, spagnolo castigliano) sono quelli pubblicati dall'*IPA* come file sonori allegati al *Handbook* o ad articoli comparsi nel *JIPA*, altri 16 (2 finlandesi, 1 cinese mandarino, 1 cinese cantonese, 1 giapponese, 2 islandesi, 2 italiani, 1 portoghese brasiliano, 1 arabo, 1 turco, 1 russo, 1 ceco, 1 tedesco, 1 romeno transilvano) sono stati registrati presso la cabina silente del Laboratorio di Fonetica Sperimentale 'Arturo Genre' di Torino, mentre altri 13 campioni sono stati registrati durante inchieste sul campo (8 islandesi, 4 varietà di romeno,[11] 1 francese canadese).[12]

I vantaggi del nostro approccio sono i seguenti:

- la segmentazione e l'etichettatura sono svolte indipendentemente dai due autori; è in genere poi calcolata la media dei risultati ottenuti con i valori misurati da entrambi ed è calcolato l'accordo tra segmentatori, al fine di garantire una migliore affidabilità;[13]
- i dati sono uniformi dal punto di vista del tipo di parlato (letto), della durata (tra i 30 e i 60 secondi per campione) e del contenuto (la stessa storia tradotta nelle varie lingue in un numero comparabile di unità prosodiche);
- l'utilizzo di *Correlatore* ha permesso di calcolare tutti i correlati di cui siamo a conoscenza (%V, ΔV, ΔC, varcoV, varcoC, rpviV, rpviC, npviV, npviC, cciV, cciC) per tutti i dati;
- il numero di lingue studiate è abbastanza alto, comunque superiore a molti altri studi di questo tipo, e in crescita.

Lo svantaggio principale consiste invece nell'aver utilizzato solo uno o due parlanti per ogni varietà linguistica presa in analisi (a eccezione dell'italiano e dell'islandese, per le quali disponiamo rispettivamente di 3 e 10 parlanti); bisogna quindi tener presente questo limite ed evitare interpretazioni assolute di quelle che potrebbero invece essere semplici idiosincrasie di un parlante.

[11] Precisiamo che si tratta di varietà di romeno letterario (olteno, munteno e moldavo), non di dialetti romeni.

[12] Ringraziamo Pier Luigi Salza per il campione di francese canadese.

[13] Per questo aspetto, che non verrà approfondito in questa sede, si veda Mairano & Romano (2007a e b).

6. I GRAFICI

In questo paragrafo presenteremo i risultati ottenuti sui nostri dati per tutti i correlati citati.[14] Per questa pubblicazione ci limitiamo a presentare i grafici ottenuti con il calcolo dei correlati secondo il metodo A,[15] che ci sembra essere il più consueto negli altri studi.

Per quanto riguarda le metriche di Ramus *et al.* (1999), i risultati ottenuti applicando queste formule sono mostrati nelle figure 8 e 9. Nel grafico della figura 8, ogni punto rappresenta un parlante, ad eccezione dell'islandese, di cui si è calcolata la media dei 10 parlanti e la deviazione standard (quest'ultima è mostrata come barra d'errore).

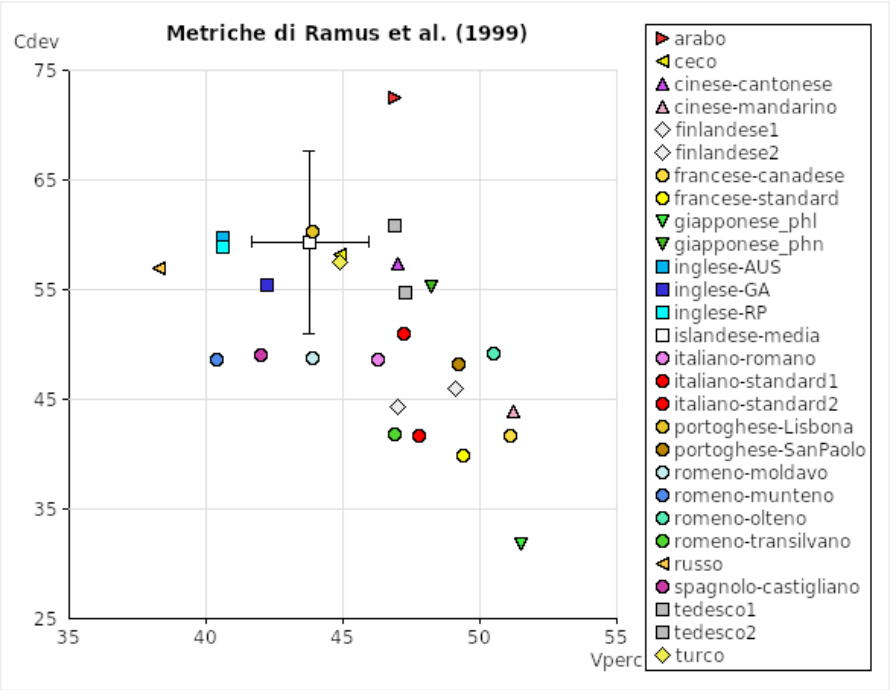


Figura 8: Grafico di ΔC vs. %V

La figura 8 mostra il grafico ΔC vs. %V, che i tre autori identificano come quello che meglio riesce a creare una distinzione tra lingue sillabiche e accentuali. In effetti, conformemente alle previsioni, si nota un gruppo di lingue tradizionalmente definite isosillabiche (italiano, francese, portoghese brasiliano, finlandese e cinese mandarino) nella parte inferiore destra del grafico, cui corrispondono valori bassi di ΔC e alti di %V.

[14] Ovviamente ci limitiamo a constatare la dispersione di misure relative a una sola produzione per un numero limitato di parlanti. La proposta di raggruppare i punti sul grafico conformemente alla suddivisione tradizionale è naturalmente arbitraria e non è suffragata da valutazioni statistiche che in questi casi sarebbero invece auspicabili.

[15] Questo significa che le metriche sono state calcolate applicando le formule sull'insieme delle misure. Si veda il paragrafo 3 per maggior dettagli.

Parallelamente, si nota un gruppo di lingue tradizionalmente classificate come isoaccentuali (inglese, tedesco, portoghese europeo, islandese, russo, arabo e cinese cantonese) nella parte superiore sinistra del grafico, cui corrispondono valori inversi di ΔC e %V. Tuttavia, stupiscono le posizioni di moldavo, munteno e spagnolo, con valori inaspettatamente bassi di %V e medi di ΔC. Per quanto riguarda il giapponese, nel grafico sono stati inseriti due valori, molto diversi tra loro ma che si riferiscono allo stesso campione etichettato in due modi diversi: giapponese_phn indica un'etichettatura di tipo prevalentemente fonetico, in cui le vocali sorde sono state considerate consonanti (con il risultato che sono venuti a formarsi cluster consonantici di tipo 'ccc' in cui la 'c' centrale è in realtà una vocale desonorizzata, il che viene rispecchiato da un ΔC molto più alto), mentre giapponese_phl indica un'etichettatura basata prevalentemente su criteri fonologici, in cui le vocali desonorizzate sono state considerate come segmenti vocalici.

Una situazione parzialmente diversa emerge invece dall'osservazione del grafico in figura 9, che mostra invece la distribuzione delle lingue per ΔC/ΔV.

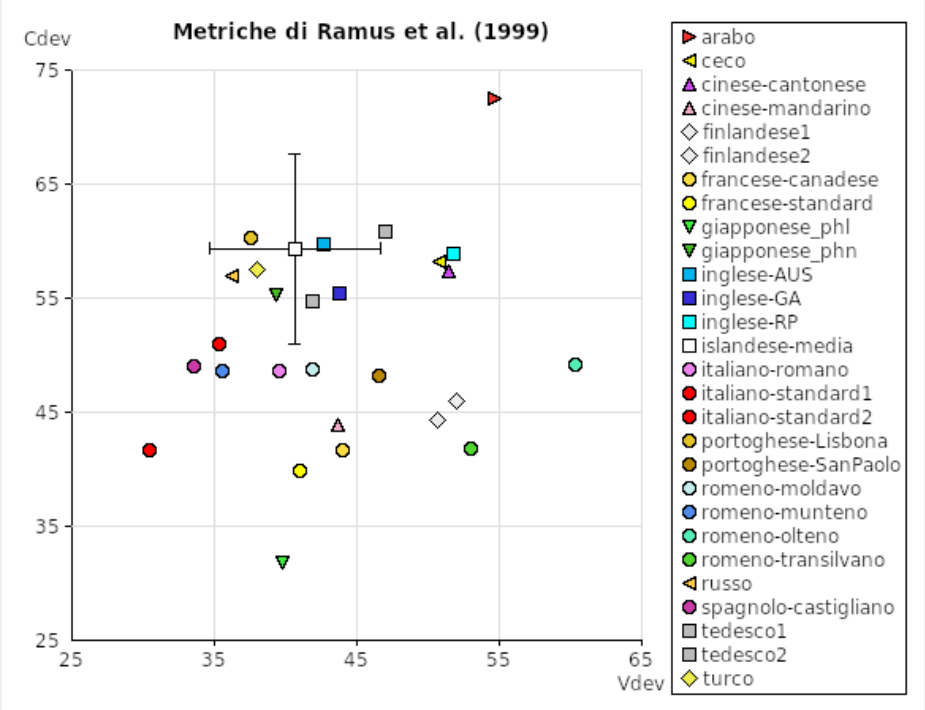


Figura 9: Grafico di ΔC vs. ΔV

Sebbene risultino ancora visibili i due gruppi tradizionali di lingue (questa volta le lingue sillabiche si trovano in basso a sinistra, mentre quelle accentuali in alto a destra), la separazione è piuttosto confusa (soprattutto in relazione a ΔV, come già notato anche da Ramus *et al.*, 1999, e da Ramus, 2002); inoltre, questa volta sono il giapponese, il finlandese e il transilvano a differenziarsi e occupare una zona particolare, con valori alti di ΔV e bassi o medio-bassi di ΔC.

Una situazione diversa si vede in figura 10, dove vengono mostrati i risultati del calcolo di VarcoV e VarcoC (ricordiamo che si ottengono calcolando la deviazione standard e dividendo i risultati per la durata media dei segmenti).

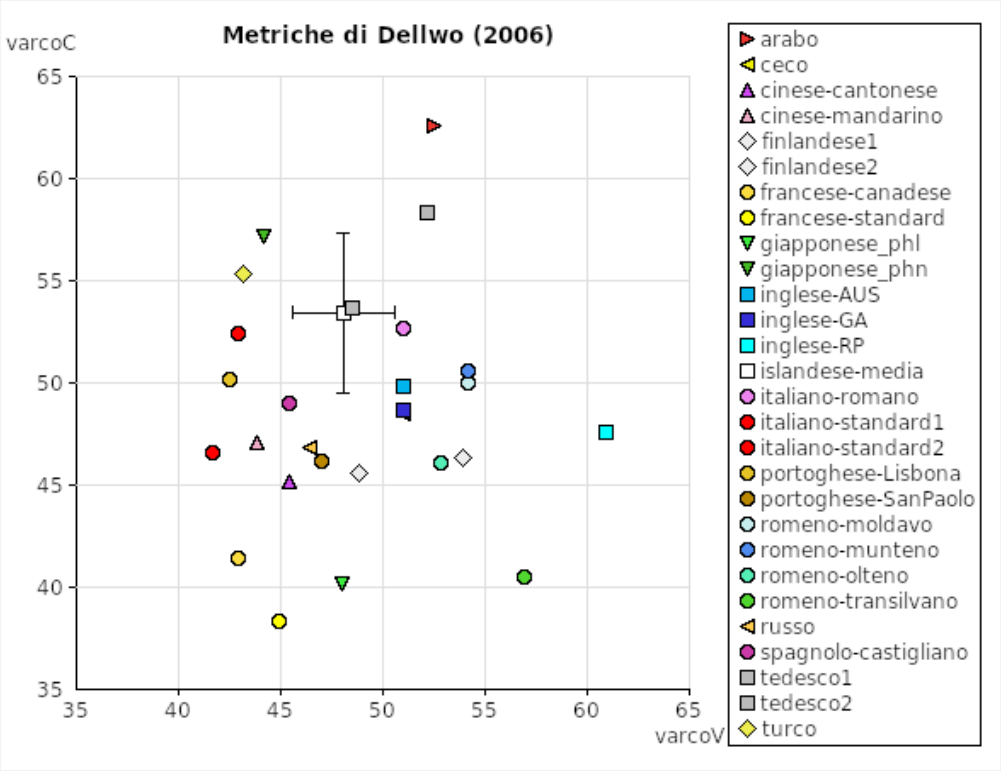


Figura 10: Grafico di VarcoC vs. VarcoV.

Anche in questo caso sarebbe possibile tracciare una linea di demarcazione tra i due gruppi di lingue, ma il giapponese e le tre varietà di romeno risulterebbero tra le lingue accentuali, mentre il finlandese si situerebbe in una posizione intermedia tra i due. Una distinzione rimane comunque difficile poiché questi due gruppi sarebbero identificati principalmente sulla base dei valori vocalici (cioè di VarcoV), mentre sarebbero più confusi considerando i valori consonantici.

Il grafico dei PVI (figura 11) mostra una concentrazione di vari campioni alla cui periferia si distinguono, agli estremi opposti, quelli di alcune lingue rappresentative dei due gruppi tradizionali. Nel quadrante in basso a sinistra, cui corrispondono valori bassi di nPVI e rPVI, troviamo distintamente i due campioni di giapponese (in funzione dei diversi criteri di etichettatura), il francese, lo spagnolo castigliano e, avvicinandosi al centro, due varietà di romeno (munteno e moldavo), un finlandese, il cinese mandarino e uno dei tre parlanti di italiano, conformemente alle aspettative tradizionali. I due portoghesi (brasiliano ed europeo), il romeno transilvano nonché un finlandese, pur presentando valori medio-bassi di rPVI, si collocano nel quadrante in basso a destra, con valori medio-alti di nPVI.

Nella porzione in alto a destra, con valori alti di rPVI e nPVI, troviamo invece l'inglese RP, l'australiano, uno dei campioni tedeschi e, avvicinadosi al centro del grafico, l'arabo, il cinese cantonese e l'altro tedesco. Due dei campioni di italiano, il romeno olteno, il ceco, il russo, il turco e l'islandese si collocano al centro del grafico, a metà strada tra i due gruppi.

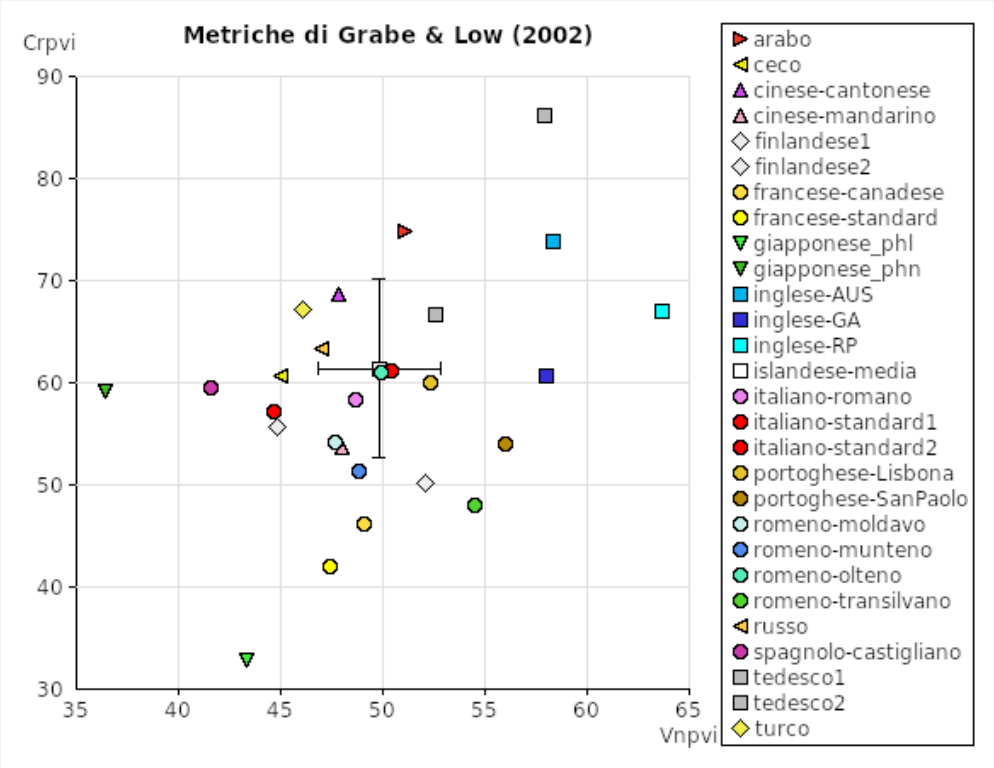


Figura 11: Grafico di rPVI (consonantico) vs. nPVI (vocalico)

Nonostante l'islandese si trovi in posizioni più accentuali nei grafici presentati in precedenza, notiamo che questa collocazione intermedia sembra per la verità abbastanza adeguata per questa lingua, in cui sono sì presenti gruppi consonantici complessi, ma in cui non vi sono fenomeni macroscopici di riduzione vocalica (infatti, a differenza di molte altre lingue germaniche, in islandese non esiste uno schwa fonologico e il timbro delle vocali atone viene mantenuto).

Passiamo infine a commentare l'ultimo grafico, in figura 12, che mostra i risultati dei CCI. In rosso è stata tracciata la bisettrice, poiché secondo le previsioni di Bertinetto & Bertini (2008) le lingue a controllo dovrebbero distribuirsi attorno a essa, mentre le lingue a compensazione al di sotto di essa. Si nota immediatamente che per alcune lingue si ottengono risultati molto netti e conformi alle aspettative; in particolare, citiamo il francese, il cinese mandarino, il romeno munteno, transilvano e moldavo, che si distribuiscono tutti attorno alla bisettrice; inoltre, il giapponese, il finlandese e anche lo spagnolo (che veniva invece collocato in una posizione più intermedia da altre metriche) sembrano quasi formare

un gruppo a parte al di sopra della bisettrice; al contrario, il tedesco, l'inglese britannico e americano e il romeno olteno si posizionano al di sotto della bisettrice, conformemente a quanto predetto per le lingue a compensazione. Lasciano invece perplessi i risultati ottenuti sul russo, l'arabo, su uno dei 2 parlanti di italiano e, in parte, sul portoghese europeo, l'inglese australiano e l'islandese che risulterebbero a controllo.

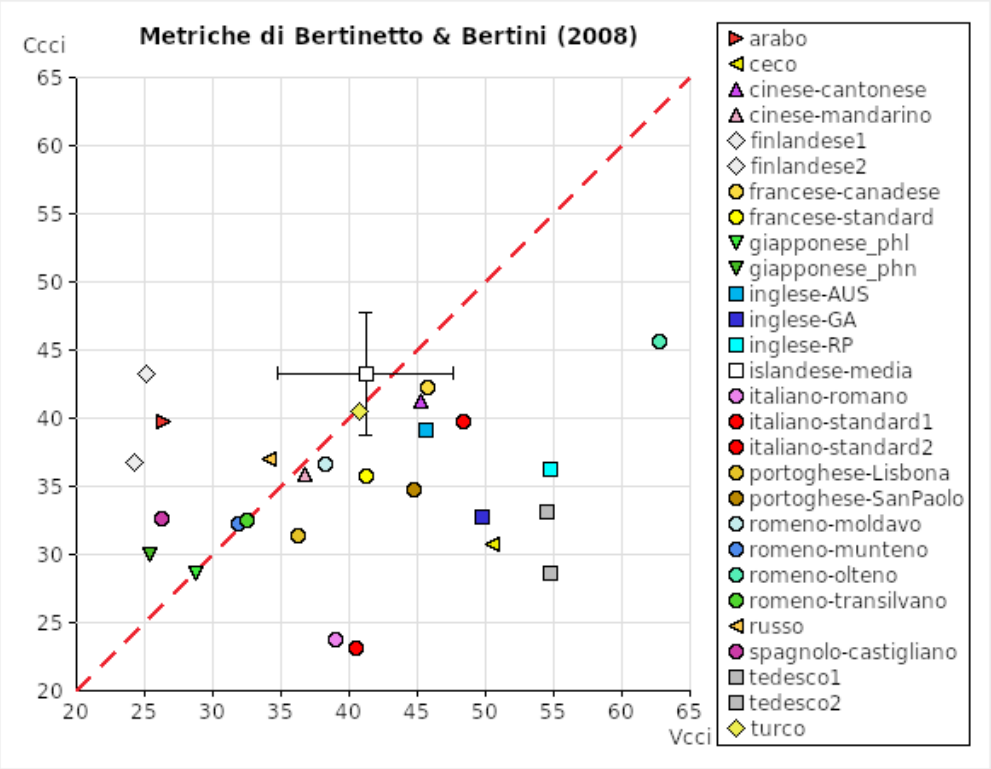


Figura 12: Grafico di Ccci vs. Ccci

7. CONCLUSIONI

Dall'analisi dei risultati ottenuti emerge chiaramente che, nonostante alcune lingue subiscano la stessa categorizzazione qualsiasi siano le metriche utilizzate (è il caso dell'inglese, del tedesco e del francese), altre lingue vengono classificate come appartenenti a gruppi diversi a seconda di quale metrica venga utilizzata (si vedano per esempio il giapponese, l'arabo, il turco e il ceco); ne consegue naturalmente che le varie metriche proposte non sono equivalenti e ognuna di esse propone una classificazione leggermente diversa delle lingue. Questo può essere naturalmente dovuto almeno in parte all'esiguità dei campioni analizzati per ogni lingua (in alcuni casi anche uno solo), che può naturalmente portare a considerare come assoluti dai dati che rispecchiano invece idiosincrasie di uno specifico parlante. Tuttavia, a noi sembra comunque molto probabile che queste differenze siano dovute al fatto che i diversi correlati colgono diversi fenomeni fonotattici delle lingue studiate.

Un problema già emerso in altri studi è che non vi è modo di dimostrare quale raggruppamento meglio descriva la realtà ritmica delle lingue. Al momento stiamo realizzando dei test percettivi sulle stesse registrazioni utilizzate per calcolare le metriche che ci permettano di stimare quanto un determinato campione linguistico suoni accentuale o sillabico nell'intenzione di ottenere un quadro di riferimento sulla base del quale dare una valutazione delle corrispondenza tra i risultati delle metriche e i dati percettivi.

Un'altra direzione di ricerca che abbiamo intrapreso recentemente si basa sulla definizione di ritmo come successione e gerarchia di prominenze al livello di piede o di sillaba.[16] Partendo da queste premesse abbiamo provato ad applicare su alcuni campioni linguistici le formule dei delta e dei PVI sui valori di intensità e frequenza fondamentale misurata in quarti di tono; in effetti, l'idea di includere questi altri due parametri nello studio del ritmo non è del tutto nuova (si veda Lee & Todd, 2004, e i risultati di una nostra prima applicazione in Mairano & Romano, 2008b). I risultati ci paiono di difficile interpretazione, ma per il futuro auspichiamo un'integrazione nello studio del ritmo linguistico di altri parametri prosodici oltre alla durata (in particolare f_0 e intensità), nella convinzione che i dati relativi a quest'ultima siano essenziali ma non sufficienti alla descrizione di questo fenomeno.

RINGRAZIAMENTI

Siamo debitori ai tre revisori anonimi che ci hanno aiutati a migliorare il nostro contributo, segnalando imperfezioni e proponendo modifiche, talvolta sostanziali, che abbiamo riconosciuto come necessarie. Eventuali ulteriori refusi sono comunque da imputare agli autori.

[16] Questa ipotesi trova riscontro, tra l'altro, nella tradizione poetica delle letterature di queste lingue, come è già stato accennato in precedenza; le lingue accentuali avrebbero, infatti, adottato un sistema metrico basato sul piede poiché gli accenti sono in queste lingue molto prominenti e quindi facilmente e intuitivamente individuabili; al contrario, le lingue sillabiche avrebbero sviluppati un sistema metrico basato sulle sillabe poiché gli accenti non sono prominenti né facilmente individuabili; sulle gerarchie d'accenti si veda tra gli altri Nespor (1993), e, più in generale, Liberman & Prince (1977). Alcuni autori sostengono anche che i parlanti di lingue sillabiche siano intuitivamente in grado di dividere una parola in sillabe, mentre i parlanti di lingue accentuali abbiano più difficoltà in questo compito, sebbene riescano meglio a individuare gli accenti di una frase. Questi indizi sembrerebbero dunque supportare l'ipotesi che gli accenti siano più prominenti nelle lingue accentuali. D'altra parte l'impressione che le sillabe accentate siano più prominenti nelle lingue accentuali che in quelle sillabiche è già codificata nei confronti 'dinetici' di Luciano Canepari (v. ad esempio Canepari, 2004: 217).

8. BIBLIOGRAFIA

Abercrombie, D. (1967), *Elements of General Phonetics*, Edimburgo: University Press.

Allen, G.D. (1975), Speech rhythm: its relation to performance universals and articulatory timing, *Journal of Phonetics*, 3, 75-86.

API – Archivio del Parlato Italiano (2001), DVD a cura di F. Albano Leoni, Università degli Studi di Napoli, *CIRASS*.

Barbosa, P. (2006), *Incursões em torno do ritmo da fala*, Campinas: Pontes.

Barry, W. & Russo, M. (2003), Isocronia Soggettiva o Oggettiva? Relazioni tra Tempo Articolatorio e Quantificazione Ritmica, in *Il Parlato Italiano* (F. Albano Leoni *et al.*, editors), Napoli: D'Auria.

Barry, W.J., Andreeva, B., Russo, M., Dimitrova, S. & Kostadinova, T. (2003), Do rhythm measures tell us anything about language type? in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, August 3-9, 2003, 2693-2696.

Bertinetto, P.M. (1977), *Syllabic Blood* ovvero l'italiano come lingua ad isocronismo sillabico, *Studi di Grammatica Italiana*, 6, 69-96.

Bertinetto (1981), *Strutture prosodiche dell'italiano. Accento, quantità, sillaba, giuntura, fondamenti metrici*, Firenze: Accademia della Crusca.

Bertinetto, P.M. (1983), Ancora sull'italiano come lingua ad isocronia sillabica, in *Scritti linguistici in onore di G.B. Pellegrini*, II, Pisa: Pacini, 1073-1082.

Bertinetto, P.M. (1989), Reflections on the dichotomy 'stress' vs. 'syllable-timing', *Revue de Phonétique Appliquée*, Mons, 99-130.

Bertinetto, P.M. (1990), Coarticolazione e ritmo nelle lingue naturali, *Rivista Italiana di Acustica*, XIV/2-3, 69-74.

Bertinetto, P.M. & Magno Caldognetto, E. (1993), Ritmo e intonazione, in *Introduzione all'italiano contemporaneo. Le strutture* (A.A. Sobrero, editor), Roma-Bari: Laterza, 141-192.

Bertinetto, P.M. & Bertini, C. (2008), On modeling the rhythm of natural languages, *Proceedings of Speech Prosody 2008*, Campinas, Brasil, May 6-9, 427-430.

Bertinetto, P.M. & Bertini, C. (in preparazione), *Towards a unified predictive model of Speech Rhythm*.

Bertini, C. & Bertinetto, P.M. (2009), Prospezioni sulla struttura ritmica dell'italiano basate sul corpus semispontaneo AVIP/API, in *La Fonetica Sperimentale. Metodo e Applicazioni* (L. Romito, V. Galatà, R. Lio, editors), Atti del 4° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Arcavacata di Rende (CS), 3-5 dicembre 2007, Torriana: EDK Editore, 3-21.

Canepari, L. (2004), *Manuale di Pronuncia*, München: Lincom.

Dankovicová, J. & Dellwo, V. (2007), Czech speech rhythm and the rhythm class hypothesis, in *Proceedings of the 16th International Congress of Phonetic Sciences,* Saarbrücken, Germany, August 6-10, 2007, 1241-1244.

Dauer, R.M. (1983), Stress-timing and syllable-timing reanalyzed, *Journal of Phonetics,* 11, 51-62.

Dellwo, V. (2006), Rhythm and speech rate: a variation coefficient for ΔC, in *Language and Language Processing: Proceedings of the 38th Linguistic Colloquium*, Piliscsaba 2003 (Karnowski, P. & Szigeti, I., editors), Frankfurt: Peter Lang, 231–241.

Dellwo, V. (2008), The role of speech rate in perceiving speech rhythm, in *Proceedings of Speech Prosody 2008*, Campinas, Brasil, May 6-9, 2008, 155-158.

Dellwo, V. & Wagner, P. (2003), Relations between language rhythm and speech rate, in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, August 3-9, 2003, 471-474.

Farnetani, E., & Kori, Sh. (1986), Effects of Syllable and Word Structure on Segmental Durations in Spoken Italian, *Speech Communication*, 5, 17-34.

Farnetani, E. & Kori, Sh. (1990), Rhythmic Structure in Italian Noun Phrases: A Study on Vowel Durations, *Phonetica*, 47, 50-65.

Fowler, C. (1977), *Timing control in speech production*, Bloomington: Indiana University Linguistic Club.

Gibbon, D. & Gut, U. (2001), Measuring speech rhythm, *Proceedings of Eurospeech 2001*, Aalborg, Denmark, September 3-7, 95-98.

Giordano, R. (2008), On the phonetics of rhythm of Italian: patterns of duration in preplanned and spontaneous speech, in *Proceedings of Speech Prosody 2008*, Campinas, Brasil, May 6-9, 74-77.

Grabe, E. & Low, E.L. (2002), Durational variability in speech and the rhythm class hypothesis, in *Papers in Laboratory Phonology* 7 (C. Gussenhoven, N. Warner, editors), Berlin: Mouton de Gruyter, 515-546.

IPA (1949), *The Principles of the International Phonetic Association*, Londra: University College (ristampa 1966).

IPA (1999), *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press (illustrazioni sonore disponibili on-line all'indirizzo: http://web.uvic.ca/ling/resources/ipa/handbook.htm).

Jian, H. (2004), On the syllable timing in Taiwan English, in *Proceedings of Speech Prosody 2004*, Nara, Japan, March 23-26, 247-250.

Lee, C.S. & McAngus Todd, N. (2004), Towards an auditory account of speech rhythm: application of a model of the auditory 'primal sketch' to two multi-language corpora, *Cognition*, 93, 225-254.

Liberman, M. & Prince, A. (1977), On Stress and Linguistic Rhythm, *Linguistic Inquiry*, 8, 249-336.

Lindblom, Bj. & Rapp K. (1973), Some temporal regularities of spoken Swedish, *Papers in Linguistics from the University of Stockholm*, 21, 1-59.

Lloyd James, A. (1940), *Speech signal in telephony*, Londra: Pitman & Sons.

Mairano, P., & Romano, A. (2007a), Lingue isosillabiche e isoaccentuali: misurazioni strumentali su campioni di italiano, francese, inglese e tedesco, in *Scienze vocali e del linguaggio. Metodologie di valutazione e risorse linguistiche* (V. Giordani, V. Bruseghini & P. Cosi, editors), Atti del 3° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, 29 novembre – 1 dicembre 2006, Povo (Trento), Torriana (RN): EDK editore, 119-134.

Mairano, P. & Romano, A. (2007b), Inter-Subject Agreement in Rhythm Evaluation for Four Languages (English, French, German, Italian), in *Proceedings of the 16th International Congress of Phonetic Sciences,* Saarbrücken, Germany, August 6-10, 2007, 1149-1152.

Mairano, P. & Romano, A. (2008a), A comparison of four rhythm metrics for six languages, Poster presentato al workshop *Empirical Approaches to Speech Rhythm* (University College London, 2008).

Mairano P. & Romano A. (2008b), Distances rythmiques entre variétés romanes, in *La variation diatopique de l'intonation dans le domaine roumain et roman*, Atti del simposio internazionale, Iaşi, Romania, 2008 (A Turculeț, editor), Iaşi: Univ. I.A. Cuza.

Marotta, G. (1985), *Modelli e misure ritmiche: la durata vocalica in italiano*, Bologna: Zanichelli.

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996), Coping with linguistic diversity: the infant's viewpoint, in *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition* (J.L. Morgan & K. Demuth, editors), Mahwah, NJ: Lawrence Erlbaum Associates, 101-116.

Mendicino, A. & Romito, L. (1991), «Isocronia» e «base di articolazione»: uno studio su alcune varietà meridionali, *Quaderni del Dipartimento di Linguistica dell'Università della Calabria*, S. L. 3, 49-67.

Mok, P.P.K. & Dellwo, V. (2008), Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English, in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, May 6-9, 2008, 63-66.

Molinu, L. & Romano, A. (1999), La syllabe dans un parler roman de l'Italie du Sud (variété salentine de Parabita – Lecce), in *Actes du Colloque des Journées d'Etudes Linguistiques:'SyllabeS',* Nantes, 25-27 mars 1999*,* Nantes, France, 148-153.

Nazzi, T., Bertoncini, J. & Mehler, J. (1998), Language Discrimination by Newborns: towards an understanding of the role of rhythm, *Journal of Experimental Psychology: Human Perception and Performance*, 24(3), 756-766.

Nespor, M. (1993), *Fonologia*, Bologna: Il Mulino.

O'Dell, M. & Nieminen, T. (1999), Coupled oscillators model of speech rhythm, in *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, USA,

August 1-7, 1999, 1075-1078.

Pamies Bertrán, A. (1999), Prosodic Typology: On the Dichotomy between *Stress-Timed* and *Syllable-Timed* Languages, *Language Design*, 2, 103-130.

Pike, K.L. (1945), *The Intonation of American English*, Ann Arbor: University of Michigan Press.

Ramus, F. (2002), Acoustic Correlates of Linguistic Rhythm: Perspectives, in *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, April 11-13, 2002, 115-120.

Ramus, F., Nespor, M. & Mehler, J. (1999), Correlates of linguistic rhythm in the speech signal, *Cognition*, 73/3, 265-292.

Roach, P. (1982), On the distinction between 'stress-timed' and 'syllable-timed' languages, in *Linguistic controversies* (D. Crystal, editor), Londra: Edward Arnold, 73-79.

Roach, P. (editor) (2003), *A Bibliography of Timing and Rhythm in Speech* (University of Reading, disponibile online all'indirizzo: www.personal.rdg.ac.uk/~llsroach/timing.pdf, ultima modifica: 2 aprile 2003).

Romano, A. (2003), Accento e intonazione in un'area di transizione del Salento centro-meridionale, in *Storia politica e storia linguistica dell'Italia meridionale*, Atti del convegno internazionale di studi parlangeliani, Messina, Italy, 2000 (P. Radici Colace, G. Falcone & A. Zumbo, editors), Messina-Napoli: Ed. Scientifiche Italiane, 169-181.

Romito, L. & Trumper, J. (1993), Problemi teorici e sperimentali posti dall'isocronia, *Quaderni del Dipartimento di Linguistica dell'Università della Calabria*, S. L. 4, 10, 89-118.

Russo, M. & Barry, W.J. (2008), Isochrony reconsidered. Objectifying relations between Rhythm Measures and Speech Tempo, in *Proceedings of Speech Prosody 2008*, Campinas, Brasil, May 6-9, 52-55.

Sachs, C. (1953). *Rhythm and tempo: a study in music history*. Londra: Dent.

Schmid, S. (1996), A typological view of syllable structure in some Italian dialects, in *Certamen Phonologicum III* (P.M. Bertinetto, L. Gaeta, G. Jetchev & D. Michaels, editors), Papers from the Third Cortona Phonology Meeting, April 1996, Torino: Rosenberg & Sellier, 247-265.

Schmid, S. (2001), Un nouveau fondement phonétique pour la typologie rythmique des langues, *Poster présenté au 10ème anniversaire du laboratoire d'analyse informatique de la parole (LAIP)*, Université de Lausanne.

Schmid, S. (2004), Une approche phonétique de l'isochronie dans quelques dialectes italo-romans, in *Nouveaux départs en phonologie* (T. Meisenburg, M. Selig, editors), Tübingen: Narr, 109-124.

Vayra, M., Avesani, C. & Fowler, C., (1984), Patterns of temporal compression in spoken Italian, in *Proceedings of the 10th International Congress of Phonetic Sciences,* Utrecht, The Netherlands, August 1-6, 1983, 2, 541-546.

White, L., Mattys, S.L., Series, L. & Gage, S. (2007), Rhythm Metrics Predict Rhythmic Discrimination, in *Proceedings of the 16th International Congress of Phonetic Sciences,* Saarbrücken, Germany, August 6-10, 2007, 1009-1012.

VARIABILITÀ RITMICA DI VARIETÀ DIALETTALI DEL PIEMONTE

Antonio Romano [a,b,c], Paolo Mairano [a,c], Barbara Pollifrone [b,c]
[a] LFSAG – Laboratorio di Fonetica Sperimentale 'Arturo Genre'
[b] Facoltà di Lingue e Letterature Straniere
[c] Università degli Studi di Torino
antonio.romano@unito.it, paolomairano@gmail.com, polbarbara@yahoo.it

1. SOMMARIO

Nel vasto e variegato panorama dei dialetti gallo-italici, le parlate del Piemonte costituiscono uno spazio tutt'altro che omogeneo (Berruto, 1974; Telmon, 1988, 2001). Tuttavia, nonostante l'attenzione prestata alle caratteristiche segmentali di queste parlate, meno studi si sono occupati degli aspetti sovrasegmentali, e in particolare ritmici. Un contributo alla collocazione ritmica delle parlate piemontesi in rapporto ad altre varietà italo-romanze è stato recentemente proposto da Schmid (2004) che ha sfruttato i dati presenti nel disco allegato a Berruto (1974) per mostrarne l'appartenenza a un gruppo di lingue dalle caratteristiche più isoaccentuali (IA) rispetto ad altre varietà italo-romanze più isosillabiche (IS).

Riferendoci allo stesso quadro metodologico, in questo studio ci proponiamo di passare in rassegna, con una tecnica sperimentale già estensivamente adottata in studi precedenti su altre lingue (v. Mairano & Romano, 2007), le caratteristiche ritmiche di alcune parlate piuttosto distanti tra loro, appositamente scelte alla periferia di questa regione linguistica. Si tratta di quelle di Roccaforte Ligure (AL), Briga Alta (CN), Exilles (TO), Capanne di Marcarolo (AL), Campertogno (VC) e Bagnolo Piemonte (CN).

I risultati collocano agli estremi opposti le due varietà liguri: quella di Roccaforte L. (fortemente caratterizzata per via dei suoi dittonghi discendenti) si situa infatti tra quelle più IA (alti ΔV e ΔC), mentre quella di Capanne di M. (che conserva meglio il vocalismo atono finale ed evita gli allungamenti in sillaba chiusa) tra quelle IS (medio ΔV e basso ΔC). Anche la varietà di Bagnolo P. si colloca in area IA in prossimità di quella di Campertogno, gravitante in area lombarda, che mostra il più alto ΔC. Exilles e Briga A. (rispettivamente di area occitana e ligure) si caratterizzano infine per un ΔC medio ma un alto ΔV. Le distinzioni rispecchiano inoltre altri fenomeni, come appunto quelli legati alle riduzioni postaccentuali.

2. INTRODUZIONE

Che le parlate del Piemonte costituiscano uno spazio tutt'altro che omogeneo costituisce un dato ben noto che è già stato oggetto di numerosi studi (si vedano in particolare Berruto, 1974; Telmon, 1988 e 2001); questa situazione viene normalmente spiegata, da un lato, con la presenza di aree d'insediamento di minoranze linguistiche storiche (di comunità pluri-lingui e stratificazioni di fenomeni areali che ne rendono difficile una delimitazione certa), dall'altro, con il fatto che le parlate di questa regione sono soggette a fenomeni di contatto con quelle delle aree contigue oppure ricadono addirittura in aree linguistiche diverse da quella piemontese (come avviene per alcune delle varietà qui considerate che sono di tipo occitano, ligure o in aree di transizione con sistemi di tipo lombardo).

Nel corso degli anni si sono susseguite indagini varie e differenziate in prospettiva geolinguistica, come i grandi cantieri atlantistici e gli studi specifici di fonetica acustica condotti su variabili diverse (tra gli altri, Genre 1980 e 1992). Questi si sono soffermati molto dettagliatamente sulle caratteristiche segmentali delle parlate, mentre gli aspetti prosodici sono stati più spesso trascurati. Con la convinzione che anche queste caratteristiche possano concorrere alla caratterizzazione delle varie parlate e contribuire a una loro distinta classificazione (anche in termini sovrasegmentali), nel presente studio, proponiamo un'indagine preliminare sulle caratteristiche ritmiche di alcune delle parlate di questa regione alla luce dei recenti metodi di tipologizzazione ritmica e in rapporto ad altri dati da noi raccolti e presentati in altri studi (si veda principalmente il contributo di Mairano & Romano, in questo stesso volume).

3. QUADRO METODOLOGICO

Rinviando per la discussione riguardo allo studio del ritmo delle lingue naturali ad altri contributi presenti in questo stesso volume (v. ad es. Romano o Mairano & Romano), concentriamo l'obiettivo della presente indagine all'applicazione di alcuni tra i più recenti metodi di valutazione ritmica basati su misure di durata a una selezione di campioni di parlato in diverse varietà. L'articolo si inserisce infatti tra quegli studi descrittivi che propongono di differenziare due o più gruppi di lingue sulla base di alcuni indici della strutturazione ritmica ricavabili da misure di durata eseguite sugli intervalli consonantici e vocalici presenti in catene di parlato.[1] Verranno utilizzate le metriche ritmiche presenti in alcuni approcci che propongono il ricorso a misure basate su un calcolo della deviazione standard o di altri indici simili.[2]

Per quanto concerne la struttura ritmica del piemontese, essa è stata oggetto di alcune indagini nel corso degli anni che hanno cercato di fornire una panoramica sulla natura ritmica delle varietà italo-romanze. In questi contributi – tra i quali citiamo quelli di Trumper *et al.* (1991) e Mayerthaler (1996), cui si riferisce Schmid (2004: 111-112), e quelli di Mendicino & Romito (1991) e Romito & Trumper (1993) – sembra emergere il fatto che le varietà piemontesi prese in considerazione (essenzialmente torinesi) presentano tratti che le avvicinano al tipo ritmico tradizionalmente denominato isoaccentuale. In anni più recenti, i contributi di Schmid (2001, 2004) hanno confermato questa ipotesi sulla base di alcune proprietà fonologiche del dialetto piemontese: sulla scia di Dauer (1983) e Bertinetto (1989), Schmid (2004) ha indagato le proprietà fonologiche e le occorrenze di diversi tipi sillabici (in base all'osservazione degli inventari sillabici) in 9 dialetti italiani (friulano, milanese, torinese, siciliano, bitontino, feltrino, napoletano, veneziano, pisano) e, successivamente, ha verificato la corrispondenza tra questi sistemi e la classificazione suggerita dagli indicatori di Ramus *et al.* (1999), riscontrando per la parlata piemontese i valori di delta più alti e di percentuale vocalica più bassi (ΔC intorno a 50 e %V a 45).

---

[1] Come noto, i primi contributi in questo senso sono di scuola anglosassone, in un ambito di ricerca inaugurato da Pike (1945) e Abercrombie (1967).

[2] Oltre che in base ai cosiddetti *Delta* (ΔC, ΔV e %V; cfr. Ramus *et al.*, 1999), abbiamo condotto le nostre valutazioni secondo i seguenti indici: i nPVI(V) e rPVI(C) basati sul *Pairwise variability Index* di Grabe & Low (2002), i VarcoV e VarcoC basati sulle formule di normalizzazione proposte da Dellwo & Wagner (2003) e i CCI(V) e CCI(C) basati sul *Control and Compensation Index* di Bertinetto & Bertini (2008).

4. I DATI

Per questo studio preliminare, abbiamo scelto parlate distanti tra loro, tutte piuttosto periferiche rispetto alla regione amministrativa del Piemonte: Roccaforte Ligure (AL), Briga Alta (CN), Exilles (TO) e Capanne di Marcarolo (AL) (studiate a partire dai dati raccolti presso il nostro laboratorio e analizzati nei volumi 27, 28, 30 e 33 dell'*Atlante Toponomastico del Piemonte Montano*; v. ATPM, 2005-2008) e di quelle di Campertogno (VC) e Bagnolo Piemonte (CN) (i cui dati sono stati raccolti rispettivamente nell'ambito della recente monografia di Molino & Romano, 2008, e della tesi di laurea inedita di Piccato, 2007); in figura 1 è mostrata la collocazione geografica di queste sei località. Benché nessuna di queste abbia beneficiato di descrizioni dialettologiche monografiche né d'inchieste specifiche, ricordiamo tuttavia le località a queste più vicine, incluse negli atlanti: Briga Marittima (ALI-94), Rochemolles (AIS-140), Gavi (AIS-169 e ALI-70), Mollia (ALI-15) e Barge (ALI-64).

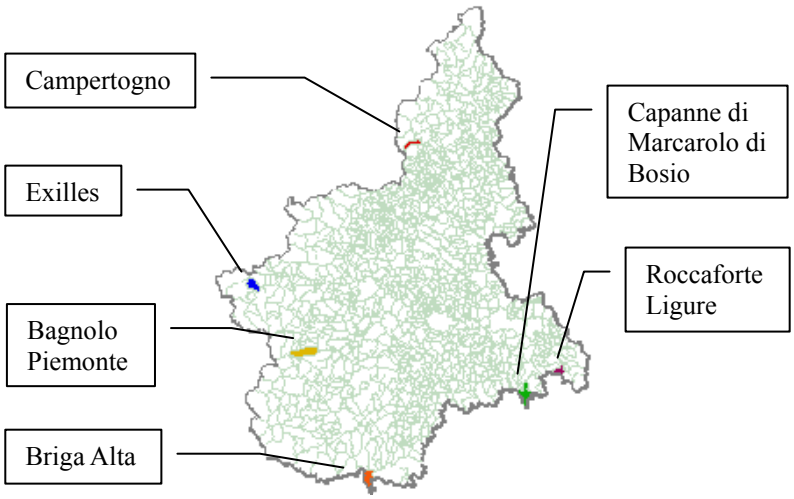


Figura 1: Carta delle suddivisione amministrative dei comuni piemontesi con l'indicazione delle località oggetto del presente studio

La tecnica adottata si basa su valutazioni effettuate su campioni di parlato letto della durata di circa 45 s (D = 34÷54 s): si tratta di versioni locali de *La tramontana e il sole* lette da un locutore per punto (tutti uomini tra i 46 e i 69 anni), le cui trascrizioni ortografiche possono essere consultate in appendice. In generale, il fatto che si sia utilizzato un solo parlante per punto implica naturalmente che i risultati non possano essere considerati totalmente rappresentativi delle varietà prese in esame. È naturalmente possibile che essi rispecchino piuttosto caratteristiche idiosincratiche dei parlanti.[3] Tuttavia, le registrazioni usate a questo scopo hanno visto coinvolti parlanti preventivamente selezionati per le raccolte di dati dell'ATPM tra quelli che la comunità rappresentata riteneva buoni

[3] Anche in quest'ambito è stato infatti dimostrato che si ottengono spesso risultati molto difformi per parlanti della stessa varietà linguistica (v. ad es. i dati di islandese in Mairano & Romano, in questo volume).

conoscitori della varietà in considerazione. In generale si tratta di cultori del patrimonio linguistico locale che hanno collaborato coi ricercatori dell'ATPM nella messa a punto dello stesso sistema di notazione ortografica usato per i testi da loro letti. Le stesse convenzioni sono state adottate dall'autrice BP nella trascrizione dei brani riportati in appendice, la cui lettura è stata fluida e spontanea.

I brani sono stati registrati nella cabina silente del Laboratorio di Fonetica Sperimentale 'Arturo Genre' di Torino e segmentati in intervalli vocalici e consonantici con *Praat*[4] secondo i criteri necessari alla successiva analisi con *Correlatore* (si veda Mairano & Romano in questo stesso volume), tramite cui sono state calcolate le metriche ritmiche – %V, ΔC, ΔV; varcoC, varcoV; rPVI(V), rPVI(V), nPVI(V), nPVI(C); CCI(V), CCI(C) – e sono stati costruiti i grafici mostrati nel paragrafo seguente. La tabella 1 riporta le strutture degli intervalli vocalici e consonantici riscontrate nelle registrazioni e le loro rispettive occorrenze, mentre la tabella 2 riassume i valori di alcune variabili globali che contraddistinguono i 6 campioni di parlato.

	Bagnolo P.	Briga A.	Camper-togno	Capanne di M. di B.	Exilles	Roccaforte L.
#	28	30	32	28	39	35
c	95	89	107	144	11	117
cc	30	50	53	33	40	34
ccc	2	4	6	2	3	6
v	110	135	147	170	156	148
vv	11	21	17	23	22	22
vvv	0	0	0	0	0	1

Tabella 1: Numero di occorrenze dei tipi d'intervallo (consonantico o vocalico) e delle pause nei 6 campioni analizzati

	Bagnolo P.	Briga A.	Camper-togno	Capanne di M.	Exilles	Rocca-forte L.
Durata senza pause	29,63 s	29,33 s	36,37 s	28,40 s	37,05 s	41,12 s
No. di sillabe	121 σ	156 σ	164 σ	193 σ	178 σ	171 σ
Velocità d'eloquio	4,08 σ/s	5,32 σ/s	4,51 σ/s	6,80 σ/s	4,80 σ/s	4,16 σ/s

Tabella 2: Valori quantitativi per alcune variabili temporali globali nei 6 campioni analizzati

[4] Le segmentazioni dei brani sono state eseguite dall'autrice BP. Etichettature e allineamenti sono stati uniformati a quelli svolti per le lingue del campione in Mairano & Romano (2007) dall'autore AR.

5. GRAFICI E DISCUSSIONE

Il grafico alla figura 2 mostra le metriche proposte da Ramus *et al.* (1999), ovvero la percentuale vocalica (%V), la deviazione standard degli intervalli consonantici (ΔC) e la deviazione standard degli intervalli vocalici (ΔV). I campioni delle sei varietà in esame sono stati integrati insieme a due campioni di francese (parigino e canadese) e di inglese (britannico e americano, più precisamente RP e GA), varietà considerate rispettivamente isosillabiche e isoaccentuali, al fine di avere dei termini di paragone.

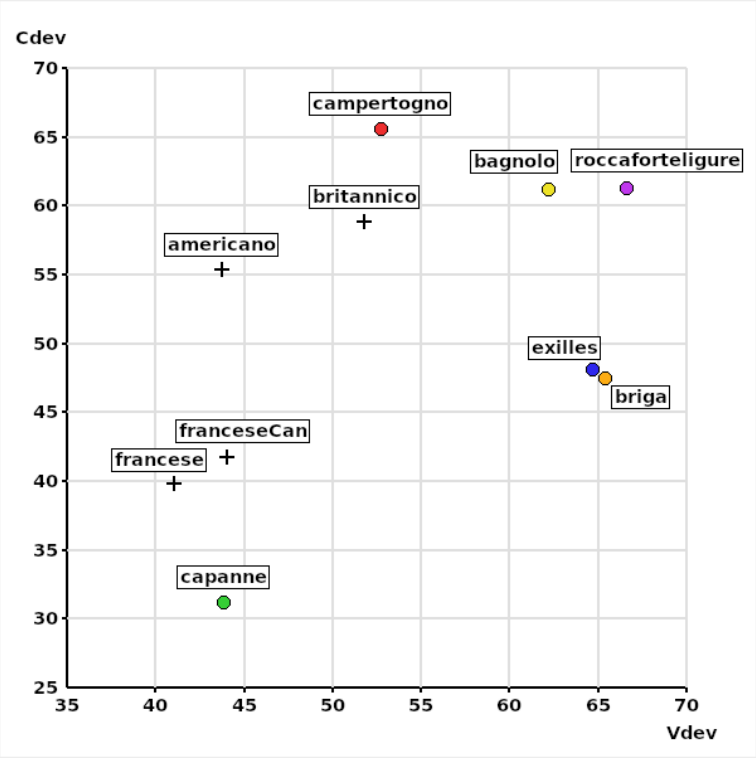


Figura 2: Grafico DeltaC vs. DeltaV per i 6 campioni analizzati a confronto con quattro lingue prese come riferimento (due varietà isoaccentuali, inglese americano e britannico, e due varietà isosillabiche, francese canadese e parigino)

I risultati collocano agli estremi opposti le due varietà di tipo ligure: quella di Roccaforte Ligure (fortemente caratterizzata per via dei suoi dittonghi discendenti) si situa infatti tra quelle più isoaccentuali (alti ΔV e ΔC), vicino alla varietà di Bagnolo Piemonte, mentre quella di Capanne di Marcarolo (che conserva meglio le vocali atone finali ed evita gli allungamenti in sillaba chiusa) tra quelle isosillabiche (ΔV medio-basso e ΔC basso). Tra l'altro, va notato che questo campione presenta la velocità d'eloquio più alta (6, 80 σ/s) e questo potrebbe confermare le ipotesi di Dellwo & Wagner (2003), secondo cui l'aumentare della velocità sposta un campione verso il polo isosillabico del continuum. La varietà di Campertogno, gravitante in area lombarda, si mostra quella col più alto ΔC, fatto che

rispecchia i risultati ottenuti da Schmid (2004), in cui è il milanese a presentare i valori più alti per questo parametro. Le varietà di Exilles e Briga Alta (rispettivamente di area occitana e ligure) si caratterizzano invece per un ΔC medio e un alto ΔV. Inoltre, le distinzioni non mancano di seguire il gradiente di altri fenomeni, come appunto quelli legati a una diversa frequenza di occorrenza di tipi fonotattici distinti (CCC, CC o V, VV) e alle riduzioni postaccentuali: se infatti a Exilles *forte* perde totalmente la sua ultima sillaba (come accade, con dati incostanti, per Bagnolo), a Campertogno si ha ancora la perdita della sola vocale finale e a Capanne la conservazione (con Briga e Roccaforte ancora propense a perderla o a desonorizzarla più spesso).

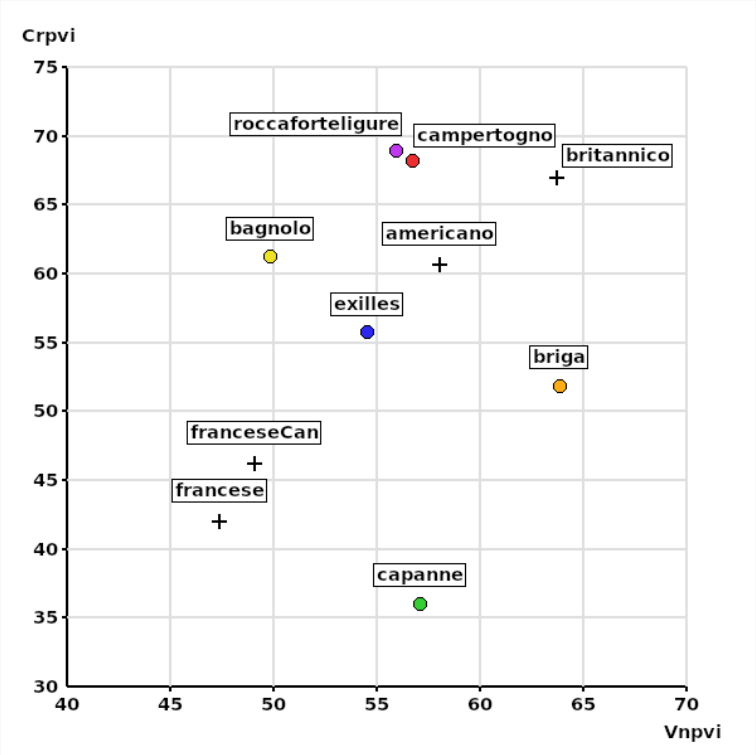


Figura 3: Grafico rPVI(C) vs. nPVI(V) per i 6 campioni analizzati a confronto con quattro lingue prese come riferimento (v. Fig. 2)

Per quanto riguarda i PVI, si può notare nella figura 3 che gli rPVI (che, secondo quanto suggerito da Grabe & Low, 2002, sono stati calcolati sugli intervalli consonantici) rispecchiano abbastanza bene i valori di ΔC, mentre gli nPVI (calcolati sugli intervalli vocalici) presentano alcune differenze rispetto ai risultati di ΔV: in particolare, Exilles, Bagnolo Piemonte e Roccaforte Ligure si sono spostati a sinistra (quindi in direzione isosillabica) pur mantenendo quasi inalterate le loro distanze reciproche; dunque, Briga Alta e Roccaforte Ligure rimangono alla stesso livello (sempre per quanto riguarda i valori vocalici), mentre Bagnolo Piemonte si trova più a sinistra. Viceversa, il campione di Capanne di Marcarolo si è spostato a destra, in direzione isoaccentuale. Si può forse

interpretare questo fatto immaginando che la normalizzazione attuata dalla formula degli nPVI abbia portato, in generale, a un appiattimento dei valori vocalici.

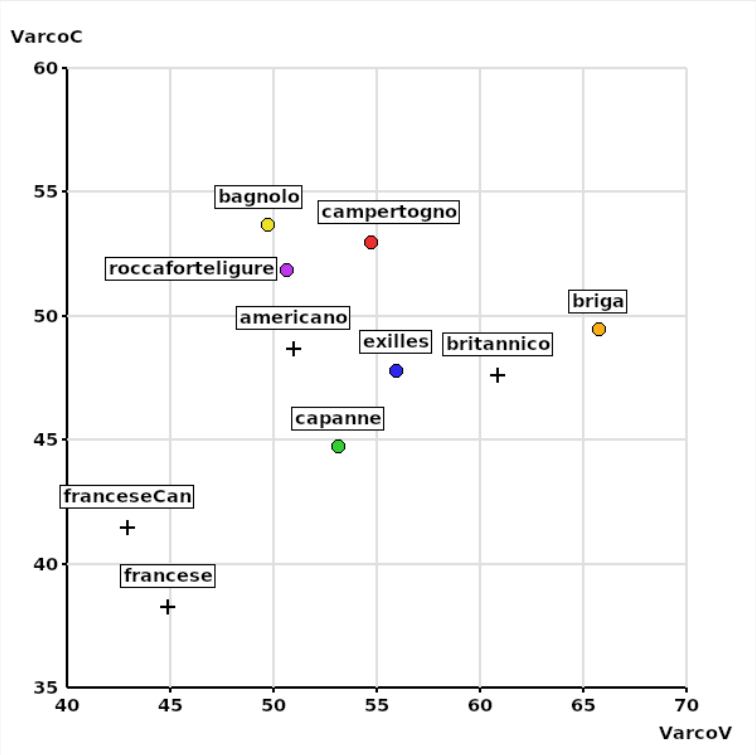


Figura 4: Grafico VarcoC vs. VarcoV per i 6 campioni analizzati a confronto con quattro lingue prese come riferimento (v. Fig. 2)

Lo stesso si evince osservando la figura 4, che mostra i valori di VarcoC e VarcoV: la normalizzazione attuata dalla formula dei Varco (questa volta sia sui valori vocalici, sia su quelli consonantici), che richiede la divisione del risultato della deviazione standard per la durata media, costituisce probabilmente il motivo per cui i vari campioni hanno subito un raggruppamento nella parte alta del grafico, con valori di VarcoC e VarcoV meno distanziati e tendenzialmente alti, quindi in area isoaccentuale.

Non rimane che analizzare il grafico dei CCI riportato nella figura 5.

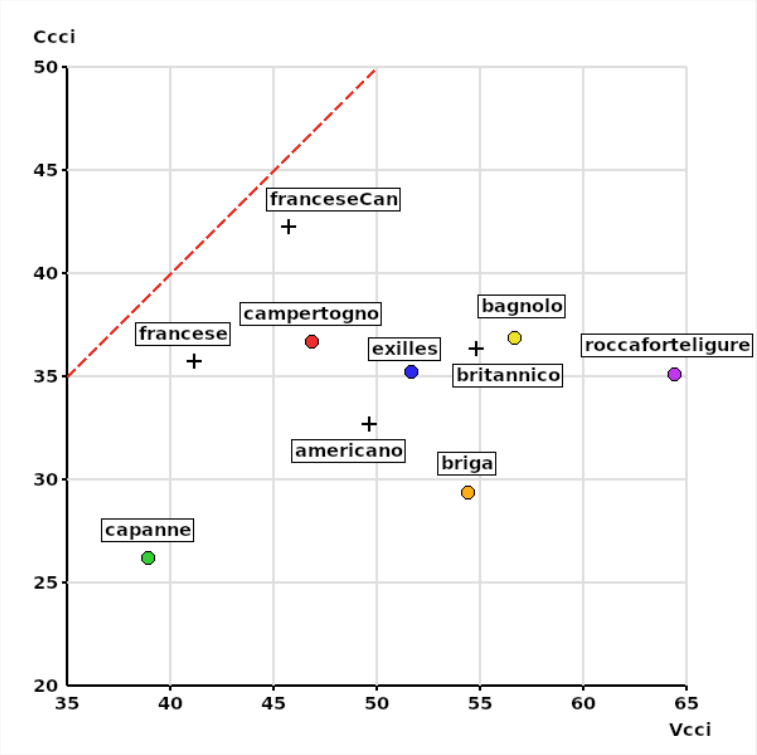


Figura 5: Grafico CCI(C) vs. CCI(V) per i 6 campioni analizzati a confronto con quattro lingue prese come riferimento (v. Fig. 2)

Si nota qui una situazione piuttosto diversa rispetto a quella presentata nei grafici precedenti. Precisiamo subito che, secondo le previsioni di Bertinetto & Bertini (2008), le lingue a controllo (come l'italiano e il francese) dovrebbero situarsi lungo la bisettrice, mentre le lingue a compensazione (come l'inglese e il tedesco) dovrebbero situarsi sotto di essa. Si può vedere che tutti i nostri campioni si situano al di sotto della bisettrice, ma che Campertogno (e, in misura minore, Capanne di Marcarolo) ne risulta comunque meno distante ed è anche molto vicino ai due campioni di francese; questo è in contrapposizione con i risultati degli altri correlati, in cui Campertogno assume una posizione marcatamente isoaccentuale. Per quanto concerne gli altri quattro campioni, i risultati sono invece concordanti con quelli dei grafici precedenti. Comunque, risulta interessante notare che il campione di Roccaforte Ligure occupa il posto più distante dalla bisettrice, il che indicherebbe che questa varietà è quella più propensa alla compensazione.

6. CONCLUSIONI

In questa rapida incursione sul tipo ritmico di alcune varietà dialettali del Piemonte, abbiamo confermato la tendenza generale all'isoaccentualità di questi dialetti, valutata tramite tutte le metriche proposte negli ultimi tempi da vari autori che si sono occupati di misure del ritmo del parlato.

Tuttavia la distribuzione dei campioni nel continuum che si definisce nei vari piani (ΔC-ΔV, nPVI(V)-rPVI(C), VarcoV-VarcoC e CCI(V)-CCI(C)) non è uniforme e alcuni di essi si aggregano o si distaccano diversamente a seconda delle metriche osservate (per es. Campertogno risulta isoaccentuale nei grafici dei Delta, dei Varco e dei PVI, ma a maggior controllo in quello dei CCI).

D'altra parte l'unico campione che resta sempre isolato al variare della rappresentazione in base alle diverse metriche, relativo alla varietà di Capanne, è anche quello contraddistinto da una velocità d'eloquio più alta rispetto agli altri; questa potrebbe essere la causa di una sua maggiore presunta isosillabicità (tuttavia ben illustrata anche dalla dominanza di sillabe CV visibile nella trascrizione ortografica in appendice).

RINGRAZIAMENTI

Ringraziamo i locutori all'origine delle registrazioni usate nel presente lavoro e Claudia Alessandri dell'ATPM (Atlante Toponomastico del Piemonte Montano) e Matteo Rivoira dell'ALI (Atlante Linguistico Italiano) per averci aiutati a contattarli e a organizzare la loro trasferta a Torino.

7. BIBLIOGRAFIA

ATPM - Atlante Toponomastico del Piemonte Montano (2005-2008: 27-Roccaforte Ligure; 28-Briga Alta, 30-Exilles e 33-Capanne di Marcarolo).

Berruto, G. (1974), *Piemonte e Valle d'Aosta*, in *Profilo dei dialetti italiani* (M. Cortelazzo, editor), 1, Pisa: Pacini.

Bertinetto, P.M. (1989), Reflections on the dichotomy 'stress' vs. 'syllable-timing', *Revue de Phonétique Appliquée*, 91-92-93, 99-130.

Bertinetto, P.M. & Bertini, C. (2008), On modeling the rhythm of natural languages, in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, May 6-9, 2008, 427-430.

Dauer, R.M. (1983), Stress-timing and syllable-timing reanalyzed, *Journal of Phonetics*, 11, 51-62.

Dellwo, V. (2008), The role of speech rate in perceiving speech rhythm, in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, May 6-9, 2008, 155-158.

Dellwo, V. & Wagner, P. (2003), Relations between language rhythm and speech rate, in *Proceedings of the 15th International Congress of Phonetics Sciences*, Barcelona, Spain, August 3-9, 2003, 471-474.

Genre, A. (1980), Le parlate occitano-alpine d'Italia, *Rivista Italiana di Dialettologia,* 4, 305-310.

Genre, A. (1992), Nasali e nasalizzate in Val Germanasca, *Rivista Italiana di Dialettologia*, 16, 181-224.

Grabe, E. & Low, E.L. (2002), Durational Variability in Speech and the Rhythm Class Hypothesis, in *Papers in Laboratory Phonology 7* (C. Gussenhoven & N. Warner, editors), Berlin: Mouton de Gruyter, 515-546.

Mairano, P. & Romano, A. (2007), Inter-Subject Agreement in Rhythm Evaluation for Four Languages (English, French, German, Italian), in *Proceedigs of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, August 6-10, 2007, 1149-1152.

Mairano, P. & Romano, A. (2019), Un confronto tra diverse metriche ritmiche usando Correlatore 1.0, in *La dimensione temporale del parlato. The temporal dimension of speech*, Atti del 5° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Zurigo, Svizzera, 4-6 febbraio 2009 (S. Schmid, M. Schwarzenbach & D. Studer, editors).

Mendicino, A. & Romito, L. (1991), «Isocronia» e «base di articolazione»: uno studio su alcune varietà meridionali, *Quaderni del Dipartimento di Linguistica dell'Università della Calabria*, S. L. 3, 49-67.

Molino, P. & Romano, A. (2008), *Il dialetto valsesiano nella media Valgrande*, Alessandria: Dell'Orso.

Piccato, E. (2007), *La parlata di Bagnolo Piemonte*. Tesi di Laurea (rel. A. Romano), Facoltà di Lingue e Letterature Straniere dell'Università degli Studi di Torino, inedita.

Ramus, F., Nespor, M. & Mehler, J. (1999), Correlates of linguistic rhythm in the speech signal, *Cognition*, 73, 265-292.

Romano, A. (in questo volume), Speech Rhythm and Timing: Structural Properties and Acoustic Correlates, in *La dimensione temporale del parlato*, Atti del 5° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Zurigo, Svizzera, 4-6 febbraio 2009 (S. Schmid, M. Schwarzenbach & D. Studer, editors).

Romito, L. & Trumper, J. (1993), Problemi teorici e sperimentali posti dall'isocronia, *Quaderni del Dipartimento di Lingue dell'Università della Calabria*, S. L. 4, 10, 89-118.

Schmid, S. (2001), Un nouveau fondement phonétique pour la typologie rythmique des langues, in *Écrits pour le 10ème anniversaire du laboratoire d'analyse informatique de la parole (LAIP)*, Lausanne (manuscrit).

Schmid, S. (2004), Une approche phonétique de l'isochronie dans quelques dialectes italo-romans, in *Nouveaux départs en phonologie* (T. Meisenburg & M. Selig, editors), Tübingen: Narr, 109-124.

Telmon, T. (1988), Areallinguistik II. Piemont, in *Lexikon der Romanistischen Linguistik* (G. Holtus, M. Metzeltin & Ch. Schmitt, editors), Vol. IV , Tübingen: Niemeyer, 469-485.

Telmon, T. (2001), Piemonte e Valle d'Aosta, in *Profili linguistici delle regioni* (A.A. Sobrero, editor), Bari: Laterza.

APPENDICE: TRASCRIZIONI ORTOGRAFICHE DEI TESTI

I testi letti dai locutori sono stati redatti da loro stessi in base alle convenzioni concordate coi ricercatori dell'ATPM. Le trascrizioni qui riportate sono state eseguite, sulla base delle stesse convenzioni, dall'autrice BP che ha annotato alcuni dettagli di esecuzione (come la sistematica caduta della consonante finale in *mantè(l)* nella produzione del locutore di Briga Alta oppure, ad es., la presenza di una nasale inattesa in *c(h')an fava* nel brano di Capanne di M.).

BAGNOLO PIEMONTE

L'òra e 'l sul descutían sü chi l'èra ël pi fòr(t) || can veim arivé 'n òm vestì d'ën mantèl | alura decìdan chë ël pi fòr(t) sarìa stait chi riüscìa a fe-ie gavé ël mantèl || L'òra per prima l'a cuminsà a sufié fòr(t) | ma pi sufiava pi l'òm se sarava ënt ël mantèl | finché l'òra lasa pèrdi || Ël sul a sua vòta taca a scaudé | e sübit òbliga l'òm a gavese ël mantèl || parèi l'òra l'è ubligà a amëtti chë 'l sul l'èra ël pi fòrt.

BRIGA ALTA (A binda e 'î sù)

Ën dì a binda e 'î sù i s' son méssi a descüttu sü chi di düi eî fusse ciü fòît(e) | cuand l'èan vist arivà 'n òm cun ün mantèl | a la uîa i an deciz chë ël ciü fòrîte sëîía vü cué di düi ch'eî fusse stait bòn a faî-ji lëvà ël mantè(l) || A binda pëî la pîima l'a cumensà a sciüscià cun tütta sa fòîsa | ma ciü la sciüsciava e ciü cuél òm së stringía ël mantè(l) || A la fin a binda a së dàita pëî vinta || Ël sù l'a cumensà a scaudà | e sübit ël caluî l'a custréntë l'òm a lëvasë el mantè(l) || cusì a binda l'a dü ërcunusciu chë ël sù l'èîa ciü fòît chë éla.

CAMPERTOGNO (Al vént e 'l sô)

An béll dì al vént e 'l sô i dispütéivu chi d'i dói a füssa al püsè fòrt | quand ch'j'iň vist rivê 'ň òmm cuň adöss 'ň mantéll || Alóra j'aň deċidü che 'l püsè fòrt | a sarìa cul ch'al füssa buň da tiréghi via al mantéll || Al vént par al primm l'è mutüssi a büfê püsè ch'al péiva | ma cupiü ch'al büféiva | cupiü l'àut a sa striňġéiva ant al mantéll | fintànt che a la fiň al vént l'è duvü désі par vint || Al sô a la sùa vòta | l'à gmaňsà a spléndi bèlli càud | e töst al calô l'à ubligà l'òmm a gavési 'l mantéll || Parê 'l vént l'è staċċ ubligà a 'rcugnüssi che 'l sô l'èra püsè fòrt che ċèll.

CAPANNE DI MARCAROLO DI BOSIO

'Na zgiurnà | u vèntu de tramuntan-a e u su descütèiven chi di duì l'èa ciù fòrte | cuandu ënt un momèntu an vistu pasà 'n òmu cu capottu indóssu || alù an decizu de véi chi di duì u l'ea u ciù fòrte e fise stetu bun a faghe levà páu primu u capottu d'indóssu || Alù u vèntu páu primu u l'a cumensau a tià fòrte | ma ciù tiava fòrte e ciù l'òmu se teniva u capottu streitu indóssu || A-a fin u vèntu u l'a ciantau lì de tià | che tanto l'a vistu che non resciva a fa n(i)inte || Alù u su la cumensau a lüzgì le | e sübitu | cu-u caldu c(h')an fava | l'òmu s'è levau u capottu d'indóssu || Cuscí u vèntu u la duvíu ricunusce che l'à pèrsu | e che u ciù fòrte u l'èa u su.

ROCCAFORTE LIGURE

’Na giurná-a | èr vèntu de tramontan-na e u su i descorivun sü chi de lu(ř) duj’(o)uvisse u ciü fórte | cuand’i vegh arivá ün | cun adòsu üna mantëléin-a || alù i decidun che u ciü fórte u saié stá | cu di i duji che ouvisse riuscéiu a fághe levá a mantëléin-a || Er ventu për primu l’a cumensá a bufá cun tütta a fórsa | ma ciü bufè ciü cuu-là u se strenzgé int’ a mantëléin-a | féin che al féin | er vèntu | u se dá pèrsu || U su l’a cumensá a splènde | e sübitu er cáadu l’a ubligá l’òmu a leváse a mantëléin-a | c(u)scì er vèntu la duvéju rëcunese che u su l’èa ciü fórte.

EXILLES (L’àuřă fřăi e ’l suřé)

Un jou l’àuřă fřăi e ’l suřé i dëscutià su qui dë lou dou foussë ’l plu fòř | can ăn véi ařibà un pasan cou l’ăvià | su laz eipala | un manté biăn săřằ | për cou l’ăvià fřăi. Alouř | i l’an desidè quë ’l (quë ’l) plu fòř | săřiă (e)tè (quèl) dë lou dou | foussë riushì ă fagă leà ’l manté || Alouř l’àuřă l’a ’ncumënsà̀ për přëmià ă së mittë ă souflà | e ’l plu souflà fòř | ’l plu ’l pasan s’ sëràvă din soun manté | finqué ă lă fin l’àuřă si douné për ganhằ || ’L suřé ăprè l’a ’ncumënsà̀ ă shoudà | e ’l plu shoudavă e ’l plu ’l pasan s’ gavavă ’l manté || păřiă l’àuřă (ou) l’è (e)tè oublijà̀ ă ricounòsë | quë ’l suřé ou l’è ’l plu fòř quë (y)è.

TEMPI E MODI DI CONSERVAZIONE DELLE *R* ITALIANE NEI *FRIGORIFERI* CLIPS

Alessandro Vietti [a], Lorenzo Spreafico [a], Antonio Romano [b]
[a] Centro di Ricerca Lingue, Libera Università di Bolzano
[b] Laboratorio di Fonetica Sperimentale 'Arturo Genre', Università degli Studi di Torino
alessandro.vietti@unibz.it, lorenzo.spreafico@unibz.it, antonio.romano@unito.it

1. SOMMARIO

In quest'articolo proponiamo un tentativo di caratterizzazione acustica di alcune realizzazioni di /r/ nell'italiano contemporaneo. Lo schema analitico qui discusso è stato applicato ai dati CLIPS (Corpora e lessici di italiano parlato e scritto) e a un campione raccolto a Bolzano con metodologia analoga. In particolare ci siamo concentrati sulle sequenze /VrV/ della parola *frigorifero*, cioè su 120 realizzazioni (di 8 parlanti per ciascuna delle 15 città del campione) e su 39 realizzazioni della stessa parola da parte di un gruppo di locutori altoatesini, parlanti nativi di italiano e/o di varietà di tedesco bavarese con competenze avanzate di italiano.

Nel panorama degli studi condotti sulle varietà d'italiano, se si escludono le osservazioni di Canepari (ad es. 1986 e 1999), le ricerche sulle modalità di realizzazione di /r/ sono al momento relativamente poco avanzate: studi acustici preliminari hanno soltanto sottolineato alcune caratteristiche salienti di rese piuttosto standard (cfr. tra gli altri Vagges *et al.*, 1978) oppure osservato dati dialettali specifici nell'ambito di studi con finalità più ampie (Sorianello, 2003; Felloni, 2006).

Mentre per altri domini linguistici l'argomento, già esplorato preliminarmente, incomincia ad essere affrontato più estesamente (cfr. ad es. Meyer-Eppler, 1959; Delattre, 1944 e 1971; Schiller, 1988; Recasens, 1991; Espy-Wilson *et al.*, 1997; Solé, 1999; Wiese, 2001; Docherty & Foulkes, 2001; Blecua Falgueras, 2001) per quello italiano non disponiamo di un quadro di riferimento completo. In Romano (2003, in prep.) sono esplorate numerose realizzazioni col metodo dei *loci* acustici e nel quadro della teoria della perturbazione. Questo riferimento può tuttavia risultare inadeguato quando si tratti di rendere conto di articolazioni multiple e di strategie di realizzazione che, bisognose di verifiche articolatorie, sfuggano a rappresentazioni certe in quest'ottica.

2. SCOPO DELLA RICERCA

Lo scopo della ricerca è quello di individuare alcuni indici acustici particolarmente robusti, tra quelli segnalati nella letteratura specialistica, per distinguere i diversi luoghi d'articolazione dei suoni /r/. Le realizzazioni di questi foni in italiano risentono – come noto – di una certa variabilità sociolinguistica che si associa in parte a fattori diatopici e diastratici, in parte a imprevedibili tendenze individuali e familiari. Il lavoro che qui presentiamo si pone ancora in termini esplorativi/dubitativi. In generale, infatti per i nostri dati, preliminarmente classificati su base impressionistica, non sono sembrati sempre affidabili gli indici segnalati in letteratura come indicatori esclusivi di specificità articolatorie, come l'abbassamento formantico di F_3 e F_4 nel caso di rese approssimanti alveolari (cfr. Ladefoged & Maddieson, 1996; Espy-Wilson *et al.*, 1997; cfr. anche Romano, 2003), o l'innalzamento di F_2 nel caso di uvularità/faringalità (v. Delattre 1971),

né i tempi e le caratteristiche delle transizioni, né le durate dei tempi d'interruzione o indebolimento (Recasens 1991, Solé 1999; cfr. Blecua Falgueras, 2008, Kouznetsov & Pamies Bertrán, 2008).

3. MATERIALI E METODO

I risultati che qui presentiamo partono da dati di italiano letto ricavati dalle liste di parole contenute nei materiali CLIPS e da quelle usate in un'indagine sull'italiano a Bolzano (cfr. Vietti & Spreafico, 2008).

Il campione analizzato è costituito dalle realizzazioni della parola *frigorifero*, così come realizzata da otto parlanti per ciascuno dei quindici punti di rilevazione previsti dal progetto (Bari, Bergamo, Cagliari, Catanzaro, Firenze, Genova, Lecce, Milano, Napoli, Palermo, Perugia, Parma, Roma, Torino, Venezia). I dati inediti, invece, sono costituiti da trentanove ripetizioni dello stesso lemma così come pronunciate da informanti nati e vissuti in Alto Adige, parlanti nativi di italiano (n=10) oppure della locale varietà di dialetto bavarese, ma con competenze avanzate di italiano (n=8). In entrambi i casi, i dati sono stati acquisiti seguendo i protocolli di raccolta CLIPS così da garantire la comparabilità dei dati. A tal fine sono stati impiegati un registratore digitale Marantz PMD660 e un microfono Behringer B-1; la campionatura è stata effettuata a 22 kHz; la digitalizzazione a 16 bit (cfr. anche Vietti & Spreafico, 2008).

La scelta di una tecnica di escussione dei dati quale quella della lettura di liste di parole è dovuta, come sempre in questi casi, alla volontà di adottare una modalità che consenta al ricercatore di mantenere il controllo sulle produzioni del locutore. Nel caso in esame si è mirato ad analizzare le realizzazioni di /r/ in contesto intervocalico, in particolare nella sequenza /ori/ dove si trovano in attacco di sillaba accentata.

Sono stati distinti e classificati i principali tipi acustici presenti nel campione, con una sommaria valutazione impressionistica, misurando diversi indici e testando statisticamente la loro conformità con le proprietà rilevate nelle valutazioni preliminari.

In particolare le realizzazioni di /r/ sono state classificate distinguendo quei casi nei quali nella rappresentazione spettrografica era presente un'interruzione di energia (conservazione) dai numerosi casi di riduzione in cui, attraverso realizzazioni approssimanti di estensione temporale variabile e caratterizzati da variazioni di energia e da transizioni formantiche assai differenziate, si arrivava in alcuni casi al dileguo di ogni traccia consonantica nel passaggio da /o/ a /i/.

Piuttosto che descrivere il suono cercando di determinare un *locus* dubbio (cfr. Öhman, 1966; Sorianello, 2003) abbiamo osservato le caratteristiche temporali della transizione, in particolare la maggiore o minore rapidità del movimento acustico (misurazione dei valori delle formanti delle vocali precedente e seguente nei punti stazionari e di transizione). I ΔF ricavati, cioè le variazioni di frequenza di formante, rapportati ai ΔT, cioè le durate dei tempi in cui si sviluppano, e integrati con le informazioni relative alla concavità o convessità delle curve, permettono infatti di avanzare una descrizione più completa delle transizioni. Accanto alla rilevazione di queste, sono state osservate la presenza e la consistenza numerica di *burst*, rumori o frizioni legati alle diverse strategie di articolazione.

Presentiamo in figura 1, uno schema riassuntivo delle variabili osservate e misurate in finestre di 200 ms attorno all'interruzione o a punto di flesso di F_2 nella transizione tra /o/ e /i/. Le misure hanno riguardato al momento essenzialmente i tempi e i modi dell'interruzione (presenza di uno, due o più *burst*, presenza di rumore durante

l'interruzione, tempo dell'interruzione, T) e le modalità di transizione di F_2. In particolare, sebbene si tratti qui soltanto di F_2, per ogni formante (F_2, F_3 e F_4) sono stati rilevati (v. fig. 1): 1) la variabile *Continuità formantica*, Ct, nel corso della vocale precedente (Ct_1) o seguente (Ct_2), con valore 1 in presenza d'interruzioni; 2) la variabile *Convessità della transizione*, Cv, prima e dopo l'interruzione (Cv_1 o Cv_2 = 0 indica andamento pressoché rettilineo, –1 andamento concavo, +1 andamento convesso); 3) le *Variazioni complessive*, B (tra F_{2offV1} e l'ultimo valore misurabile prima dell'interruzione) e D (tra il primo valore misurabile dopo l'interruzione e F_{2offV2}); 4) le durate delle transizioni prima (A) e dopo l'interruzione (C). Vista la rilevanza dei contributi delle altre formanti (F_3 e F_4), già segnalato da Delattre (1944), i risultati ottenuti al momento, sulla base dei soli valori relativi a F_2, sono necessariamente parziali.

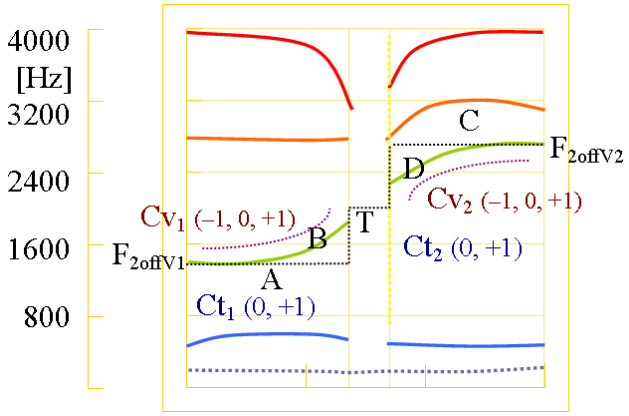


Figura 1: Schema simbolico delle variabili osservate e/o misurate

4. ANALISI DEI DATI

In base alle misure effettuate, organizzate in fogli elettronici, abbiamo proceduto a un *clustering* dei dati e a una prima valutazione quantitativa anche su base diatopica, distinguendo tuttavia realizzazioni maschili e femminili.

Possiamo confermare la distribuzione areale già illustrata nelle descrizioni più tradizionali. Troviamo quindi realizzazioni uvularizzate (velarizzate o faringalizzate) a Parma (5) e Torino (4) e *flap* (talvolta lateralizzati) a Venezia (5% complessivo sui dati nazionali).

In generale, si rileva una realizzazione dominante monovibrante 'rigida' (apico-alveolare, nel 38%; ben esemplificata nei dati di Palermo) che, rispetto a quella riportata nei dati di varietà iberiche (Recasens, 1991; Solé, 1999; Blecua Falgueras, 2001, 2008), si caratterizza per una certa 'rigidità' energetica, prima e/o dopo, che la fanno percepire (seppur non polivibrante) come più forte di una normale monovibrante (riconosciuta in un 6% di casi).[1] In un residuo 5% si presenta, invece, una monovibrante più morbida, con profili energetici più sfumati.

Altre varianti osservate:

[1] Su quest'argomento si veda ora anche Kouznetsov & Pamies Bertrán (2008).

- monovibranti di durata significativa (31±6 ms; 7%) con caratteristiche acustiche simili a quella di un'occlusiva sonora (una breve /d/ alveolare o postalveolare);
- realizzazioni approssimanti interrotte da localizzati cali di energia (18,3%);
- rese approssimanti pure (in luoghi d'articolazione diversi, 6,7% dei casi);
- realizzazioni velari, uvulari e faringali (uvularizzate o faringalizzate), approssimanti o costrittive, compaiono a Genova (1), Parma (5), Torino (4), Milano (1) e Cagliari (1) per un complessivo 10% (5% non approssimanti: l'unica chiara costrittiva uvulare è di Parma, mentre sono più comuni monovibranti o approssimanti alveolari uvularizzate; cfr. Canepari, 1999);
- una realizzazione vibratile (*flap*, talvolta lateralizzato) è infine dominante nei dati di Venezia (per un residuo 6% complessivo sui dati nazionali), rendendo i *frigoriferi* di questa località gli unici del corpus la cui provenienza geografica sia facilmente riconoscibile;
- forme di rotacismo vocalico (1,7%; con esempi isolati, da Napoli a Bergamo);
- casi di presunta cancellazione (3,3%).

Riportiamo in figura 2, la sovrapposizione di tutte le realizzazioni per i principali tipi di /r/ nel campione. In particolare i grafici della prima riga si riferiscono a realizzazioni monovibranti (in genere con una o entrambe le pareti rigide, cioè caratterizzate da un locale aumento di energia che si manifesta come un *burst* breve e intenso, qui rappresentato da un tratto discontinuo più spesso e giallo e, nelle trascrizioni, dal simbolo |; v. anche Tabella 1).[2]

Nella seconda riga di grafici, sono proposti invece gli schemi sovrapposti di tutte quelle realizzazioni che, pur classificate come approssimanti (data la continuità delle transizioni formantiche), presentano una locale (breve) caduta energetica assimilabile a un'interruzione.

[2] La ragione per cui, pur essendo presenti i due cosiddetti 'irrigidimenti', continuiamo a classificare questi suoni come monovibranti sono legate al fatto che sullo spettrogramma il numero d'interruzioni (che rivela il numero di contatti avvenuti tra l'organo mobile e l'organo fisso) si presenta pari a uno (cfr. Recasens, 1991). La discussione resta però aperta (si veda il recente contributo di Kouznetsov & Pamies Bertrán, 2008). Tra gli allofoni di questo tipo, potremmo distinguere ancora, quelli con palatalizzazione, che manifestano in generale con un maggiore abbassamento di F_1 e un aumento di F_2 anticipati prima dell'interruzione, ma spesso riconoscibile anche per una maggiore convergenza di F_2 e F_3 nella prima parte della seconda vocale (cfr. schemi in Romano, 2003). Nel nostro caso, data la difficoltà oggettiva nel discriminarli dagli altri, la distinzione è stata solo virtuale: come si vede infatti dagli schemi, sono stati trattati congiuntamente con gli altri allofoni monovibranti.

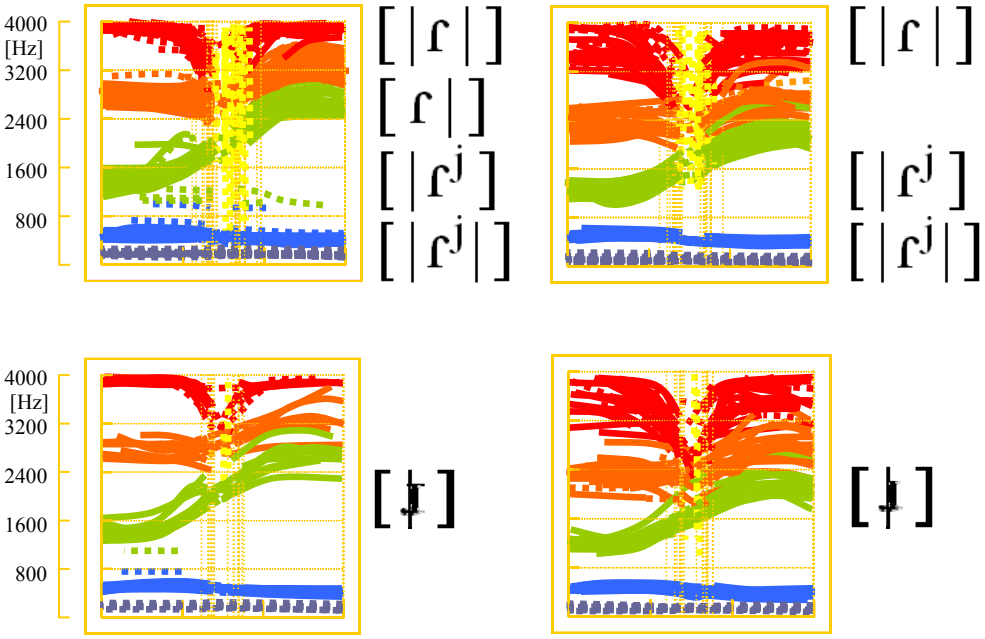


Figura 2: Schema riassuntivo dei principali tipi osservati
(voci femminili a sinistra e maschili a destra)

La fig. 2 illustra in alto le realizzazioni monovibranti ‘rigide’ apico-alveolari (l’38% dei casi; la durata dell’interruzione è pari a 24±5 ms e 27±7 ms, rispettivamente) e in basso le realizzazioni approssimanti ‘interrotte’ (18,3% dei casi; la è pari a 19±7 ms e 20±6 ms).

		n	*A* [ms]	*B* [Hz]	*B* (%)	*T* [ms]	*C* [ms]	*D* [Hz]	*D* (%)
M	*media*	16	75	479	32,2	24	68	594	23,1
	dev.st.		11	117	3,4	5	10	138	1,8
F	*media*	21	62	532	42,0	27	52	339	16,0
	dev.st.		10	100	5,5	7	13	117	5,6

Tabella 1: Dati medi sulle realizzazioni monovibranti nel corpus CLIPS

La tab. 1 riporta dati medi sulle realizzazioni monovibranti nel corpus CLIPS. Rammentiamo che A rappresenta la durata della transizione prima dell’interruzione, mentre B rappresenta l’escursione di questa. T è la durata dell’interruzione; C rappresenta invece la durata della transizione dopo l’interruzione, così come D ne rappresenta l’escursione.

Figura 3: Schema riassuntivo di tutti i tipi osservati (indistintamente dal tipo di voce) per le 15 località del corpus CLIPS

In figura 3 abbiamo invece riassunto cumulativamente gli schemi di tutte le realizzazioni osservate, raggruppate per località. Dai grafici risulta immediatamente la presenza di forme con palatalizzazione o retroflessione (visibili nei comportamenti anomali di F_2, in verde, prima dell'interruzione, maggiormente concentrati a Bari, Firenze e Venezia). Si nota invece più frequentemente nei dati di Torino un abbassamento precoce di F_2, indice di velarizzazione o uvularizzazione. Quanto a F_1, talvolta disturbata da una formante spuria di nasalità (in alcuni casi di Firenze, ad esempio), si noterà invece come un suo aumento ascendente prima della transizione (associato a un passaggio per valori maggiori di 500 Hz, prima di ridiscendere dopo l'interruzione) sia segnale inequivocabile di faringalizzazione (v. Delattre, 1944; Romano, 2003), fatto che si verifica in maniera evidente soprattutto nei dati di Parma (cfr. Felloni, 2006).

5. IL CASO DI BOLZANO

5.1 Descrizione di /r/ in Alto Adige

L'italiano parlato in Alto Adige presenta caratteristiche peculiari sia se rapportate alle dinamiche del repertorio sociolinguistico tipico del resto d'Italia – caratterizzato come noto da diglossia o dilalìa tra standard (regionale) e dialetto – sia se messe in relazione con altre comunità alloglotte storiche presenti sul territorio nazionale.

Il termine 'italiano' definisce dunque comunità di parlanti, tipi di repertori e varietà di lingua ben diverse per dinamiche costitutive, funzioni sociali ma, soprattutto, elementi e caratteristiche grammaticali. In tal senso è possibile individuare almeno tre varietà[3] situate lungo un *continuum*: anzitutto l'italiano regionale bolzanino (d'ora in avanti *STI-i*); quindi l'italiano di tedescofoni (*STI-d*); infine l'italiano di bilingui (*STI-z*).

Per via delle vicende storiche legate all'insediamento italiano nella regione altoatesina la prima varietà considerata, lo *STI-i*, risente almeno in parte dell'apporto dei dialetti italo-romanzi e, tuttavia, mantiene nel complesso le fattezze caratteristiche dell'italiano regionale di nord-est, seppur con un maggiore apporto dei dialetti veneti meridionali, in particolare per quanto riguarda le varietà diastraticamente meno alte.

La seconda varietà di italiano (*STI-d*) è più complessa da definire in termini sistemici ed è propria dei parlanti che hanno appreso il dialetto tirolese come L1 nel corso della socializzazione primaria e successivamente – e con esiti più o meno positivi in funzione dell'intensità e della continuità di accesso all'*input* – l'italiano regionale. Lo *STI-d* si caratterizza pertanto al suo interno come *continuum* di varietà di apprendimento che risentono, in modo più o meno marcato, dell'influenza del dialetto tirolese e del tedesco regionale.

La terza varietà di italiano (*STI-z*) è ugualmente associata a una classe di parlanti, in questo caso i bilingui, ovvero quelli che, grosso modo, hanno acquisito entrambi i codici durante la socializzazione primaria. Allo stato attuale mancano descrizioni linguistiche di tale varietà che, tuttavia, pur risultando in linea di massima molto simile allo *STI-i*, potrebbe risentire degli effetti dell'interazione con il dialetto tirolese.

Se, in generale, le descrizioni linguistiche delle varietà di italiano parlato in Alto Adige sono piuttosto limitate (rilevanti eccezioni si hanno però in Egger, 1979; Kramer, 1981; Coletti *et al.*, 1992; Mioni, 2001), ancora più sporadiche sono le trattazioni dedicate alla fonetica e alla fonologia di tali varietà. Di estremo interesse in tal senso sono soprattutto Mioni (1990a; 2001), Canepari (1999), Tonelli (2002) che, per quanto riguarda la distribuzione degli allofoni di /r/ in *STI-i* e *STI-d*, riportano quadri parzialmente sovrapponibili.

Mioni (1990a) afferma, sulla base dell'analisi di interazioni semi-spontanee e di liste di parole lette, che i parlanti di *STI-i* utilizzano quasi esclusivamente monovibranti alveolari [ɾ], soprattutto in contesto intervocalico. Lo stesso è confermato da Tonelli (2002: 50), che tuttavia segnala anche la comparsa di realizzazioni polivibranti, seppur solo nel caso di *hervorhebende Sprechweise*. La netta prevalenza di realizzazioni monovibranti – in linea con la tendenza nazionale (Romano, 2003; ma cfr. anche §3) – potrebbe essere dovuta anche alla forte influenza esercitata dai dialetti trentino-veneti parlati nelle aree circostanti

[3] Per una più completa panoramica sui tipi di repertori linguistici cfr. Mioni (1990b).

in conseguenza della massiccia migrazione dal Veneto meridionale registrata durante gli anni del fascismo e il dopoguerra. Peraltro la stessa sarebbe anche alla base dell'importazione di varianti più marcate regionalmente quali il *tap* retroflesso [ɽ], oppure l'approssimante retroflessa [ɻ] da noi individuate nel corso dell'analisi (cfr. *infra*) e che, apparentemente, si riscontrano soprattutto nelle varietà diastraticamente più basse.[4]

Sulla base di osservazioni impressionistiche Canepari (1999: 392) afferma invece che nella pronuncia altoatesina dell'italiano si riscontrano almeno tre distinte realizzazioni uvulari di /r/: la fricativa [ʁ], la vibrante [ʀ] e l'approssimante [ʁ̞]. L'autore sottolinea poi la possibilità di riconoscere anche un fono vibrante caratterizzato da un'articolazione complessa alveo-uvulare trascritta come [ɾ].

Vista la peculiare situazione di contatto linguistico che caratterizza l'area, anche le osservazioni sulle varianti impiegate in STI-d meritano di essere riportate. A tal proposito Mioni (1990a: 203) afferma che "gli informanti usano *tutti*[5] e categoricamente un qualche tipo di *r* uvulare". In tal caso la loro presenza è esplicitamente ricondotta all'influsso del sostrato, vale a dire dei dialetti bavaresi parlati nell'area indagata. In tal senso è tuttavia importante sottolineare che, come già osservato in Mioni (2001: 69) e Tonelli (2002) e come confermato dall'analisi delle carte del *Tirolischer Sprachatlas*, diverse varietà di dialetto altoatesino presentano realizzazioni apico-alveolari ('trillate') di /r/.

Sebbene la presenza di ognuna delle realizzazioni individuate in Mioni (1990a), Canepari (1999), Tonelli (2002) sia stata recentemente confermata (Vietti & Spreafico, 2008), va tuttavia notato come l'analisi spettro acustica di un più ampio campione di parlato spontaneo e di laboratorio (Spreafico & Vietti, in stampa) permetta di individuare anche altre varianti quali, oltre alle già citate approssimante retroflessa [ɻ] e *tap* retroflesso [ɽ], l'approssimante alveolare [ɹ], l'approssimante labiodentale [ʋ], l'approssimante uvulare [ʁ̞] oppure ancora articolazioni complesse alveo-uvulari del tipo già prospettato da Canepari (1999), che necessitano però di ulteriori analisi.

[4] Per la distinzione tra *tap* e *flap*, che necessita ancora di contributi articolatori chiarificatori, valgono qui le considerazioni già in Ladefoged & Maddieson (1996: 230-232). Cfr. tuttavia Romano (2003, in prep.).

[5] Corsivo dell'autore.

5.2 Analisi dei dati del campione di Bolzano

Il confronto dei dati *CLIPS* con il campione di italiano bolzanino fornisce, da un lato, conferme sul piano descrittivo alle tendenze nazionali e, dall'altro, permette, in ragione delle già menzionate peculiarità storico-linguistiche, di saggiare i descrittori proposti su una gamma di realizzazioni alloglotte dei suoni /r/, realizzate nella parte posteriore della cavità orale non sfruttando come articolatori primari le regioni post-dorsali e, in parte, radicali.

Una prima osservazione riguarda i tipi di foni rinvenuti nel *corpus* di 39 occorrenze di *frigorifero*. La moda del campione nazionale, la monovibrante alveolare, è confermata nei dati dei parlanti italofoni[6] (21 *tap*, più un caso realizzato come *flap*) e in più, accanto a essa, compaiono più rare realizzazioni approssimanti e una retroflessa. Anche per le varietà di italiano di tedescofoni la monovibrante uvulare, poco attestata e descritta in letteratura, risulta il fono più frequente a discapito dell'attesa polivibrante uvulare.

Sulla base delle occorrenze ridotte le ipotesi che si possono formulare sono di tipo esplorativo e impressionistico e andrebbero naturalmente suffragate da basi di dati più ampie. Ciò nondimeno, osservando la Tabella 2 su tutte le occorrenze e ancor di più la Tabella 3 sui soli *tap*, l'aspetto più rilevante sembra essere l'intervallo tra i due valori di F_2 nella fase di transizione Vr. Nel campione di locutrici in Tab. 2 il valore dell'intervallo è in termini assoluti di circa 400 Hz per le apico-alveolari e di 130 Hz per le uvulari e rispettivamente di 34 e di 10,5 in termini percentuali. Nei rispettivi grafici delle figg. 4 e 5, si può notare immediatamente come la curva di F_2 di /o/ mostri diversi *pattern* di pendenza in relazione a diversi gradi di coarticolazione con il *tap* apico-alveolare e come nel caso del *tap* uvulare la curva presenti una pendenza minima o nulla.

L'eventuale presenza di un effetto coarticolatorio è evidente (come già in Sorianello, 2003) quando sono opposti nella sequenza VC(V) suoni anteriori e posteriori: nel caso indagato, infatti, la coarticolazione anticipatoria ha luogo quando la consonante è una monovibrante alveolare e non quando a seguire si ha un *tap* uvulare. Osservando il grafico di fig. 4 e la Tab. 3 (in particolare per il sotto-campione femminile) possiamo notare come vi sia una notevole variabilità nel percorso di transizione (valore di F_2 alla fine della transizione e intervallo di F_2) nel caso del *tap* alveolare. Una possibile interpretazione del fenomeno è legata alla presenza di un *continuum* di maggiore o minore tensione ed energia articolatoria, già evidenziato per il quadro nazionale, caratterizzato da suoni che potrebbero opporre un diverso grado di resistenza alla coarticolazione (v. Recasens & Pallarès, 1999). Andrebbe pertanto suddiviso ulteriormente il campione in sotto-insiemi più omogenei al loro interno.

[6] Si intende qui i parlanti che hanno appreso l'italiano nella socializzazione primaria e solo successivamente il tedesco (*Hochdeutsch*) prevalentemente in contesto scolastico.

				A ms	B Hz	B %	T ms	C ms	D Hz	D %
M	*anteriori*	*n=11*	*media*	38	261	22,2	25	36	399	23,9
			dev.st.	10	95	8,4	11	11	99	6,9
	posteriori	*n=1*	*media*	0	0	0,0	38	42	523	34,9
			dev.st.	/	/	/	/	/	/	/
F	*anteriori*	*n=17*	*media*	46	414	34,0	27	29	421	22,2
			dev.st.	11	147	12,5	6	16	264	15,6
	posteriori	*n=10*	*media*	18	133	10,5	39	40	489	27,7
			dev.st.	21	182	13,6	16	17	292	19,8

Tabella 2: Dati medi su tutte le realizzazioni nel corpus di Bolzano (v. Tabella 1)

				A ms	B Hz	B %	T ms	C ms	D Hz	D %
anteriori	*F*	*n=13*	*media*	48	440	36,4	26	29	421	21,9
			dev.st.		36	11,1			238	
	M	*n=8*	*media*	38	283	24,1	26	37	402	23,9
			dev.st.		24	8,5			122	
posteriori	*F*	*n=5*	*media*	6	14	1,1	37	39	470	28,0
			dev.st.		1	2,5			376	

Tabella 3: Dati medi sulle realizzazioni monovibranti nel corpus di Bolzano (v. Tabella 1)

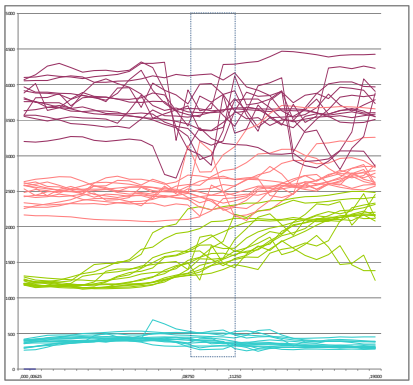


Figura 4: Tracciati formantici di F1, F2, F3 e F4 per le realizzazioni alveolari

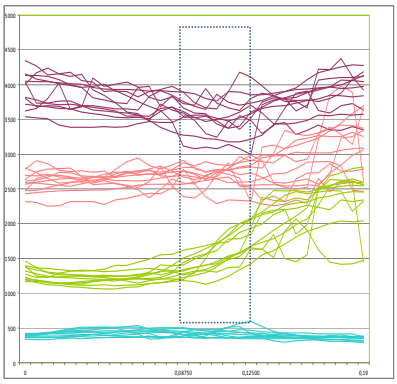


Figura 5: Tracciati formantici di F1, F2, F3 e F4 per le realizzazioni uvulari

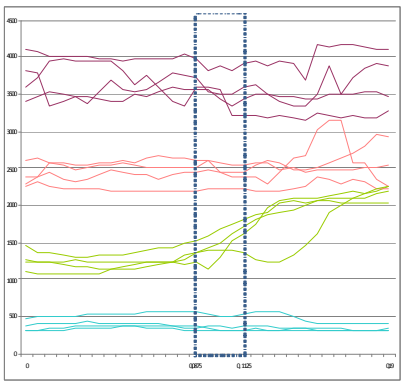


Figura 6: Tracciati formantici di F1, F2, F3 e F4 per le realizzazioni approssimanti alveolari

Un ultimo appunto interessante riguarda la transizione CV nei *tap* (Tab. 3) alveolari e uvulari dove invece non appaiono differenze rilevanti in relazione al diverso luogo di articolazione: sia i valori di F_2 (~2000 Hz) sia quelli dell'intervallo di transizione di F_2 sono simili in termini assoluti (421÷470 Hz) e divergono di poco in percentuale (21,9÷28), semmai sono i tempi della transizione a divergere con 29 ms nel caso delle monovibranti alveolari e di 39 ms nel caso di quelle uvulari (Tab. 3 campione femminile).

6. DISCUSSIONE

Allo stato attuale è stato effettuato solo un confronto dei dati complessivi relativi alle realizzazioni monovibranti alveolari. A una prima sommaria osservazione dei dati relativi a F_2 di tutte le realizzazioni presenti nel corpus *CLIPS* e di tutte quelle del corpus bolzanino, emerge immediatamente una chiara convergenza sui tempi medi d'interruzione, ma una significativa differenza relativa ai tempi delle transizioni.

Mentre non si presentano significativamente distinti gli scarti frequenziali delle transizioni sulla vocale seguente (che sembrano addirittura scambiati nei valori medi tra voci maschili e femminili), sono invece risultate sensibilmente significative le differenze tra le misure di durata delle transizioni su entrambe le vocali nel caso dei locutori maschili ($t(A)=2,49$, $p<0,02$; $t(C)=2,15$, $p<0,05$): argomento che, anche considerata la bassa probabilità di separazione statistica tra i dati maschili e femminili di questa località, lascia pensare a ragioni legate a una maggiore apertura della vocale precedente che giustifica una riduzione nei valori di B e, in modo correlato, nei tempi in cui si verifica la transizione.

Risulta quindi difficile scorporare i dati, a causa dei pesanti condizionamenti che, più di altre consonanti, le realizzazioni di /r/ subiscono dai suoni circostanti (cfr. Öhman, 1966; Sorianello, 2003): anche le valutazioni sulle variazioni percentuali risentono di queste condizioni e, soprattutto per F_2, rendono scarsamente utilizzabili le variabili A e C da noi definite. Molto meno sensibili a questo tipo di variazione si sono invece presentate le variabili B e D (soprattutto B per dati relativi alla stessa comunità linguistica): ad es., per Bolzano, un buon discrimine medio tra realizzazioni anteriori e posteriori è possibile proprio basandosi su questo parametro ($t(B)=11,83$, $p<0,001$). Ovviamente però, vista la rilevanza dei contributi delle altre formanti (F_3 e F_4), segnalata sin da Delattre (1944), è invece su queste che occorrerà concentrare le ricerche per affinare le possibilità di distinzione tra diverse realizzazioni in luoghi d'articolazione prossimi o multipli.

RINGRAZIAMENTI

Ringraziamo i parlanti bolzanini che hanno accettato di contribuire con le loro produzioni linguistiche a questa ricerca preliminare. Desideriamo inoltre ringraziare i tre revisori anonimi che, con le loro interessanti osservazioni, hanno dato un contributo decisivo al migliormento della versione finale di quest'articolo i cui limiti restano naturalmente imputabili solo ai suoi autori.

7. BIBLIOGRAFIA

CLIPS = Corpora e Lessici dell'Italiano Parlato e Scritto: http://www.clips.unina.it/

Blecua Falgueras, B. (2001), *Las vibrantes del español: manifestaciones acústicas y procesos fonéticos*, Tesi di Dottorato, Università Autonoma di Barcellona, http://www.tdx.cat/TDX-0111102-110913.

Blecua Falgueras, B. (2008), Los sonidos vibrantes: aspectos comunes y variación, in *New Trends in Experimental Phonetics* (A. Pamies & E. Melguizo, editors), IV Congreso Internacional de Fonética Experimental, Granada, Spain, February 23-25, 2008, *Language Design*, special issue 1, 23-30.

Canepari, L. (1986), Italiano standard e pronunce regionali, Padova: CLEUP.

Canepari, L. (1999), *MaPI. Manuale di Pronuncia Italiana*, Bologna: Zanichelli.

Coletti, V., Cordin, P. & Zamboni, A. (1992), Il Trentino e l'Alto Adige, in *L'italiano nelle regioni* (F. Bruni, editor), Torino: UTET.

Delattre, P. (1944), A contribution to the history of «R grasseyé», *Modern Language Notes*, December 1944, 562-564 (ripubblicato in 1966, *Studies in French and Comparative Phonetics*. *Selected papers in French and English*, The Hague: Mouton, 206-207).

Delattre, P. (1971), Pharyngeal features in the consonants of Arabic, German, Spanish, French and American English, *Phonetica*, 54, 93-108.

Docherty, G. & Foulkes, P. (2001), Variability in /r/ production. Instrumental perspectives, *Etudes et Travaux*, 4, 173-184.

Egger, K. (1979), Morphologische und syntaktische Interferenzen an der deutsch-italienischen Sprachgrenze in Südtirol, in *Standardsprache und Dialekte in mehrsprachigen Gebieten Europas* (S. Ureland, editor), Tübingen: Niemeyer, 55-104.

Espy-Wilson, C.Y., Narayanan, S., Boyce, S.E. & Alwan, A. (1997), Acoustic modelling of American English /r/, in *Proceedings of Eurospeech '97*, Rhodes, Greece, September 22-25, 393-396.

Felloni, M.C. (2006), Un'indagine sociofonetica a Parma: la realizzazione del fonema /r/ nell'italiano regionale, *Tesi di Laurea Specialistica*, Università di Pavia.

Kouznetsov, V. & Pamies Bertrán, A. (2008), Trill with one closure. Still a trill or a tap? Data from Russian and Spanish, in *New Trends in Experimental Phonetics* (A. Pamies & E. Melguizo, editors), Actas del IV Congreso Internacional de Fonética Experimental, Granada, Spain, February 23-25, 2008, *Language Design*, special issue 1, 149-160.

Kramer, J. (1981), *Deutsch und Italienisch in Südtirol*, Heidelberg: Winter.

Ladefoged, P. & Maddieson, I. (1996), *The sounds of the world's languages*, Oxford: Blackwell.

Meyer-Eppler, W. (1959), Zur Spektralstruktur der /r/-Allophone des Deutschen, *Akustica*, 9, 246-250.

Mioni, A. (1990a), La standardizzazione fonetico-fonologica a Padova e Bolzano (stile di lettura), in *L'italiano regionale* (M. A. Cortelazzo, A. Mioni, editors), Roma: Bulzoni, 193-208.

Mioni, A. (1990b), Bilinguismo intra- e intercomunitario in Alto Adige/Südtirol: considerazioni sociolinguistiche, in *Mehr als eine Sprache. Zu einer Sprachstrategie in Südtirol – Più di una lingua. Per un progetto linguistico in Alto Adige* (F. Lanthaler, editor), Merano: Alpha & Beta, 13-36.

Mioni, A. (2001), L'italiano nelle tre comunità linguistiche tirolesi, in *Die Deutsche Sprache in Südtirol* (K. Egger & F. Lanthaler, editors), Wien: Folio, 65-76.

Öhman, S.E.G. (1966), Coarticulation in VCV utterances: Spectrographic measurements, *Journal of the American Society of Acoustics*, 39, no. 1, 151-168.

Recasens, D. (1991), On the production characteristics of apicoalveolar taps and trills, *Journal of Phonetics*, 19, 267-280.

Recasens, D. & Pallarès, M. (1999), A study of /r/ and /rr/ in the light of the 'DAC' coarticulation model, *Journal of Phonetics*, 27, 143-170.

Romano, A. (2003), *A contribution to the study of phonetic variation of /r/ in French and Italian linguistic domains,* Poster presented at the 2nd International workshop on the sociolinguistic, phonetic and phonological characteristics of /r/, Université Libre de Bruxelles, 5-7 December, 2002 (in c. di p. in H. Van de Velde, R. van Hout, D. Demolin, editors, preprint 62 pp.)
http://www.personalweb.unito.it/antonio.romano/r_romano_2006.pdf .

Schiller, N. (1988), The phonetic variation of German /r/, in *Variation und Stabilität in der Wortstruktur* (M. Butt, N. Fuhrhop, editors), Hildesheim: Olms, 261-287.

Solé, M.J. (1999), Production requirements of apical trills and assimilatory behavior, in *Proceedings of the 14th International Congress of Phonetic Sciences,* San Francisco, USA, August 1-7, 1999, 487-489.

Sorianello, P. (2003), Aspetti coarticolatori nel parlato di Siena, in *La coarticolazione* (G. Marotta, editor), Atti delle Giornate del XIII Gruppo di Fonetica Sperimentale, Pisa, 28-30 novembre 2002, Pisa: ETS, 101-110.

Spreafico, L. & Vietti, A. (in stampa), Sistemi fonetici in contatto: la variabilità di /r/ nell'italiano in Alto Adige, in *La comunicazione parlata*, Atti del terzo convegno internazionale sulla comunicazione parlata, Napoli, 23-25 febbraio 2009).

Tonelli, L. (2002), *Regionale Umgangssprachen*, Padova: Unipress.

Vagges, K., Ferrero, F.E., Magno Caldognetto, E. & Lavagnoli, C. (1978), Some acoustic characteristics of Italian consonants, *Journal of Italian Linguistics*, 3, 69-85 (paper presented at the *8th International Congress of Phonetic Sciences*, Leeds 1975, preprint 23 pp.).

Vietti, A. & Spreafico, L. (2008), Phonetic variation of */r/* in a language contact context: The case of South Tyrol Italian, *Poster presented at Laboratory Phonology 11 – Phonetic detail in the lexicon*, Wellington, New Zealand, June 30-July 2, 2008.

Wiese, R. (2001), The unity and variation of German /r/, *Etudes et Travaux*, 4, 11-26.

NOTE SULLE OPPOSIZIONI DI QUANTITÀ VOCALICA

Arianna Uguzzoni
Alma Mater Studiorum – Università di Bologna
arianna.uguzzoni@unibo.it

1. SOMMARIO

Queste note affrontano in modo succinto e certamente non esaustivo il tema della caratterizzazione e classificazione di alcune lingue europee che fanno un uso distintivo delle differenze temporali.

Nella prima nota si fa una separazione tra lingue con sistema quantitativo doppio, che riguarda sia vocali sia consonanti, e lingue con sistema quantitativo unico. Per queste ultime vengono delineati, con l'appoggio di esempi, tre scenari tipologicamente diversi tra di loro, a seconda che le opposizioni di quantità abbiano il loro ambito nella sequenza /vocale+consonante/, nella /consonante/, nella /vocale/.

Argomento della seconda nota è la considerazione dei principali fattori che determinano l'organizzazione delle lingue in merito alla utilizzazione libera o vincolata delle distinzioni di quantità. Sono illustrate le condizioni fonotattiche e prosodiche entro le quali operano le opposizioni di quantità vocalica e/o consonantica: l'accento lessicale, la forma della sillaba, la posizione nella parola.

Riguardo alla indipendenza vs. dipendenza dall'accento lessicale si osserva che la distinzione fonologica tra vocali brevi e vocali lunghe nella maggior parte delle lingue europee odierne è limitata alle sillabe accentate, sia in area germanica, sia in area romanza. Per quel che concerne la distribuzione della quantità vocalica nella sillaba e nella parola sono esaminate tre situazioni: sillaba aperta all'interno di parola, sillaba chiusa, sillaba aperta in fine di parola. Una chiara differenziazione tipologica emerge dal diverso comportamento delle lingue nella prima e nella terza condizione.

La quarta nota presenta e discute aspetti della teoria che interpreta le opposizioni di quantità vocalica di alcune lingue germaniche come opposizioni prosodiche di taglio sillabico (*Silbenschnittgegensätze*). Secondo alcuni il taglio sillabico porta a una netta dicotomia fra due tipi di lingue: le *Quantitätensprachen* e le *Silbenschnittsprachen*; secondo altri esso fornisce un metodo per descrivere adeguatamente modalità di differenziazione temporale che caratterizzano un tipo 'speciale' di sistema quantitativo.

Vengono di solito considerate lingue con taglio sillabico il tedesco (precisamente il tedesco settentrionale), l'olandese, l'inglese, in cui si oppongono due modalità di interazione tra la vocale accentata e la consonante successiva: *scharfer* vs. *sanfter Schnitt.* In alcune versioni della teoria brevità e lunghezza delle vocali accentate assumono il ruolo di concomitanti fonetici dei due modi di taglio sillabico, rispettivamente, del taglio brusco e del taglio piano.

2. PRIMA NOTA: SISTEMA QUANTITATIVO DOPPIO E SISTEMA QUANTITATIVO UNICO

2.1.0 Lingue con sistema quantitativo doppio

Una prima linea demarcativa può essere tracciata tra lingue che presentano un sistema quantitativo 'doppio' e lingue che invece presentano un sistema quantitativo 'unico'.

Il finlandese possiede chiaramente un sistema quantitativo doppio, in quanto l'utilizzazione distintiva delle differenze temporali riguarda sia la /vocale/ sia la /consonante/. Le opposizioni di quantità vocalica e le opposizioni di quantità consonantica sono indipendenti le une dalle altre (Lehtonen, 1970; Suomi, 2005).

Una situazione simile si trova in ungherese, in latino, in lingue germaniche antiche. Negli esempi addotti si osservano quattro tipi di struttura di parola e quattro tipi di sequenza /vocale+consonante/. Le parole sono citate alla maniera tradizionale come ho trovato nelle fonti e ad esse non ho fatto seguire la trascrizione in IPA.

Finlandese:		
muta	ˈCVCV	V+C
muuta	ˈCV:CV	V:+C
mutta	ˈCVC:V	V+C:
muutta	ˈCV:C:V	V:+C:
Latino:		
mălus	ˈCVCV	V+C
mālus	ˈCV:CV	V:+C
mălleus	ˈCVC:V	V+C:
mālle	ˈCV:C:V	V:+C:
Antico svedese:		
gata	ˈCVCV	V+C
dööma	ˈCV:CV	V:+C
falla	ˈCVC:V	V+C:
dootter	ˈCV:C:V	V:+C:

Nel caso delle lingue germaniche mi sono limitata all'antico svedese e a parole bisillabiche, rinviando a Riad (1995) e a Becker (1998) che trattano e illustrano il tema in modo esauriente.

2.2.0 Lingue con sistema quantitativo unico

Il sistema quantitativo unico, che predomina nelle lingue europee odierne, in molti casi è da considerare il prodotto di una serie di cambiamenti che hanno fatto abbandonare un sistema quantitativo originariamente doppio. Si pensi ai processi evolutivi, attestati o ricostruiti, che hanno portato, da una parte, dal latino alle lingue romanze, dall'altra, dal germanico antico alle lingue germaniche odierne

Senza alcuna pretesa di completezza, ritengo utile delineare tre scenari che sono tipologicamente diversi tra di loro per quel che riguarda l'uso distintivo delle differenze temporali. Il modo di presentazione degli esempi farà emergere anche le forme 'legali' e le forme 'non legali' (precedute da asterisco) dal punto di vista della struttura della parola e della sequenza /vocale+consonante/.

2.2.1 Scenario I. Opposizioni di quantità nell'ambito della sequenza /vocale+consonante/

Svedese, norvegese, islandese settentrionale, bavarese centrale sono lingue accomunate da una caratteristica significativa: in una sillaba accentata o è lunga la vocale (V:+C) o è lunga la consonante postvocalica (V+C:), come mostrano le seguenti parole bisillabiche e monosillabiche.

Svedese, norvegese:

		*ˈCVCV		*CVC	*V+C
veeka	[ˈve:ka]	ˈCV:CV	viis [vi:s]	CV:C	V:+C
vekka	[ˈvek:a]	ˈCVC:V	viss [vis:]	CVC	V+C:
		*ˈCV:C:V		*CV:C .	*V:+C:

La 'complementarità quantitativa' rappresentata da /vocale lunga+consonante breve/ e da /vocale breve+consonante lunga/ ha sollevato la questione del 'primato' fonologico della quantità vocalica o della quantità consonantica. Vari argomenti sono stati portati a favore dell'una (per es. Linell, 1978) e dell'altra soluzione (per es. Eliasson, 1978). Dato che segmenti brevi e segmenti lunghi, sia vocalici sia consonantici, non entrano in opposizione diretta tra di loro, è possibile una terza interpretazione secondo cui, nelle lingue citate, il 'campo' delle opposizioni di quantità non è né la /vocale/ né la /consonante/, bensì la sequenza /vocale+consonante/ considerata come un tutto (per es. Bannert, 1976; Bannert, 1977).

2.2.2 Scenario II. Opposizioni di quantità nell'ambito della /consonante/

Le sillabe accentate di parole bisillabiche dell'italiano standard presentano una distribuzione dei fenomeni temporali analoga a quella dello scenario I. Una vocale lunga è seguita da consonante breve (V:+C), mentre una vocale breve è seguita da consonante lunga (V+C:).

Italiano:

		*ˈCVCV	*V+C
fato	[ˈfa:to]	ˈCV:CV	V:+C
fatto	[ˈfat:o]	ˈCVC:V	V+C:
		*ˈCV:C:V	*V:+C:

Nell'interpretazione di dati come questi non ci sono state controversie tra gli studiosi, concordi nell'attribuire il ruolo 'dominante' alla opposizione di quantità consonantica. La natura lunga o breve della vocale in ˈCV:CV da un lato e in ˈCVC:V dall'altro viene considerato un fatto predicibile in base allo status, rispettivamente, breve o lungo della consonante postonica. Pertanto la rappresentazione fonologica di coppie come *fato* e *fatto* è /ˈfato/ e /ˈfat:o/. Si ricorderà che l'opposizione tra consonante breve e consonante lunga in italiano si trova anche dopo vocale atona: per esempio in *camino* (/kaˈmino/) vs. *cammino* (/kaˈm:ino/). Non entro qui nel dibattito che divide sostenitori della soluzione monofonematica e sostenitori della soluzione bifonematica: in una parola come *fatto* si tratta di una consonante lunga (/ˈfat:o/) o invece di una consonante geminata (/ˈfatto/)? (Muljačić, 1972; Bertinetto, 1981; Hurch & Tonelli, 1982).

2.2.3 Scenario III. Opposizioni di quantità nell'ambito della /vocale/

Più numerose sono le lingue caratterizzate dalla presenza indiscussa di opposizioni di quantità vocalica: ceco, franco-provenzale, italo-romanzo settentrionale, friulano, danese, olandese, tedesco settentrionale. A parte il caso del ceco, l'opposizione tra vocali brevi e vocali lunghe è limitata alle sillabe accentate. In questa condizione le sequenze /vocale+consonante/ sono o del tipo V+C o del tipo V:+C. A scopo illustrativo, porto gli esempi del frignanese, una varietà di italo-romanzo settentrionale, e del tedesco settentrionale.

Frignanese:

it. 'botte'	[ˈbɔta]	ˈCVCV	it. 'pile'	[pel]	CVC	V+C
it. 'botta'	[ˈbɔ:ta]	ˈCV:CV	it. 'pelo'	[pe:l]	CV:C	V:+C
		*ˈCVC:V			*CVC:	*V+C:
		*ˈCV:C:V			*CV:C:	*V:+C:

Tedesco settentrionale:

Mitte	['mɪtə]	ˈCVCV	*Bett*	[bɛt]	CVC	V+C
Miete	['mi:tə]	ˈCV:CV	*Beet*	[be:t]	CV:C	V:+C
		*ˈCVC:V			*CVC:	*V+C:
		*ˈCV:C:V			*CV:C:	*V:+C:

È utile, a mio parere, soffermare l'attenzione sulle strutture di parola ammesse e sulle strutture di parola escluse nello scenario III a confronto con gli scenari I e II. Sono strutture 'legali' (1.) ˈCVC(V), (2.) ˈCV:C(V), mentre sono 'non legali' (3.) *ˈCVC:(V), (4.) *ˈCV:C:(V). La situazione attuale ha le radici in processi avvenuti in fasi precedenti della storia delle lingue. In questa sede è opportuno accennare almeno al fenomeno dell'abbreviamento delle consonanti lunghe, più noto con il nome di degeminazione, che si è verificato in alcune parti dell'area romanza e dell'area germanica. Nel campo, per esempio, dei tipi di sequenza /vocale+consonante/ e dei tipi di struttura di parola, la divergenza che si osserva chiaramente tra lingue come lo svedese e il norvegese, da un lato, e lingue come il danese e il tedesco settentrionale, dall'altro, è il riflesso sincronico del fatto che il processo della degeminazione delle consonanti in posizione postonica è avvenuto nelle seconde, ma non nelle prime.

Più in generale, per la ricostruzione del formarsi delle opposizioni di quantità vocalica rimando agli importanti lavori di Árnason (1980), per l'area germanica, e di Loporcaro (2005, 2007), per l'area romanza.

Al fine di rendere più evidente la distinzione tipologica fra III e I, nella trascrizione fonetica delle parole ho di proposito omesso la segnalazione dell'allungamento allofonico che si riscontra acusticamente nella consonante postonica collocata dopo vocale breve. Tale allungamento è assente in danese; mentre si manifesta nelle altre lingue del gruppo III in modo variabile da lingua a lingua e da soggetto a soggetto. Merita di essere tenuta presente una sistematica differenza di comportamento che si constata all'interno di qualcuna di queste lingue a seconda che si tratti di bisillabi o di monosillabi (per es. olandese, frignanese, bolognese). Una trattazione più particolareggiata del fenomeno dell'allungamento consonantico si può vedere in Uguzzoni & Busà (1995).

Per mostrare in quale misura e in quali condizioni la brevità della vocale accentata sia accompagnata da una maggiore estensione temporale della consonante che la segue, faccio riferimento ad alcuni dati riguardanti il tedesco settentrionale e il frignanese. Da una ricerca

di Fischer-Jørgensen & Jørgensen (1969) risulta che nel tedesco settentrionale, sia in parole bisillabiche sia in parole monosillabiche**,** in media le consonanti successive a vocale breve sono più lunghe del 27% rispetto a quelle successive a vocale lunga (il rapporto tra le due durate è pari a 78,74). Varie ricerche sul frignanese indicano con chiarezza che soltanto in posizione finale (cioè in parole monosillabiche) la durata acustica della consonante collocata dopo vocale breve (CVC) è maggiore di quella collocata dopo vocale lunga (CV:C): in media nel primo caso la consonante postonica è più lunga del 23% rispetto al secondo caso (il rapporto tra le due durate è pari a 81,19). In posizione interna, invece (cioè in parole bisillabiche), si riscontrano differenze fisiche trascurabili tra la consonante successiva a vocale breve (ˈCVCV) e la consonante successiva a vocale lunga (ˈCV:CV): in media nel primo caso la consonante postonica è più lunga soltanto dello 0,1% rispetto al secondo caso; il rapporto tra le due durate è pari a 98,57 (Uguzzoni & Busà, 1995; Uguzzoni, 2006a, 2006b: 120).

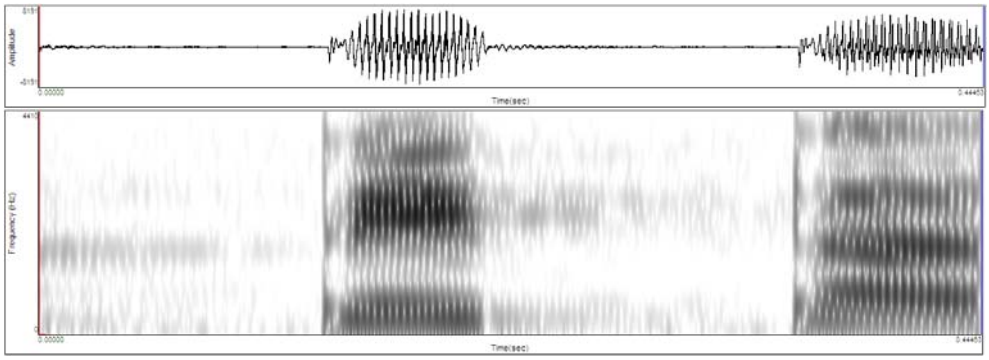


Figura 1: Forma d'onda e spettrogramma della parola [ˈpepa]
(durata totale 444 ms: vocale tonica 61 ms, consonante postonica 161 ms)
{audio 1}

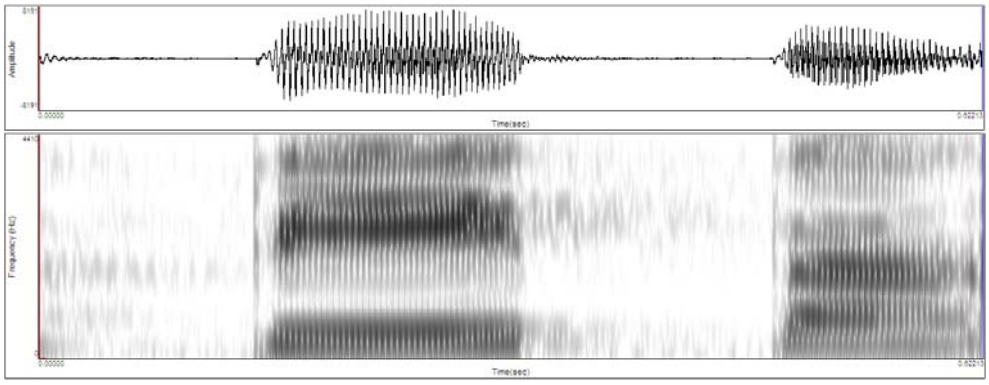


Figura 2: Forma d'onda e spettrogramma della parola [ˈpe:pa]
(durata totale 621 ms: vocale tonica 157 ms, consonante postonica 177 ms)
{audio 2}

3. SECONDA NOTA: DIFFERENZE FONOTATTICHE E PROSODICHE TRASVERSALI AI TIPI QUANTITATIVI

3.1.0 Condizioni fonotattiche e prosodiche

Un secondo criterio usato nella individuazione di tipi e sottotipi è costituito dalle condizioni fonotattiche e prosodiche entro le quali operano le opposizioni di quantità. Anche da questo punto di vista le opposizioni di quantità vocalica e/o consonantica non mostrano un quadro unitario, ma sono caratterizzate da differenze che 'attraversano' le lingue di cui si è proposta una classificazione, sia pure provvisoria, nelle pagine precedenti. Tali differenze saranno l'argomento di questa nota, in cui parleremo in modo succinto di alcuni casi di condizionamento esercitato dall'accento, dalla forma della sillaba, dalla posizione nella parola.

3.1.1 Indipendenza vs. dipendenza delle opposizioni di quantità dall'accento lessicale

Un'importante biforcazione tipologica ha il suo fondamento nel ruolo dell'accento lessicale. Le relazioni di quantità (brevità vs. lunghezza) e le relazioni di accento (sillaba accentata vs. non accentata) possono avere due modalità distinte: essere indipendenti tra di loro o essere dipendenti tra di loro. Nelle lingue considerate in questo articolo le opposizioni di quantità mostrano, con regolarità, o indipendenza (A) o dipendenza (B) dalla presenza dell'accento lessicale.

Vediamo il caso (A). In finlandese, in ungherese, in latino e nelle fasi antiche delle lingue germaniche (che appartengono al gruppo visto in 2.1.0), le opposizioni di quantità vocalica e di quantità consonantica sono indipendenti dall'accento lessicale, nel senso che ricorrono tanto in sillabe accentate quanto in sillabe non accentate. In finlandese per esempio vocali brevi si oppongono a vocali lunghe sia nella prima sillaba di parole come [ˈt–k–], [ˈt–:k–], [ˈt–k:–], [ˈt–:k:–]: la ratio V/V: è pari a 50.00 (1:2.00), sia nella seconda sillaba di parole come [ˈt–k–], [ˈt–k–:], [ˈt–:k–:], [ˈt–:k:–:]: la ratio V/V: è pari a 72.46 (1:1.38). Il finlandese e l'antico alto tedesco ammettono, da una parte, vocali lunghe in una sillaba non accentata (per esempio finlandese [ˈt–:k–:]), dall'altra, consonanti lunghe dopo una vocale non accentata (per esempio finlandese [ˈ–:te:l:inen]).

Dal punto di vista dell'indipendenza delle opposizioni di quantità vocalica dall'accento lessicale, il ceco (che appartiene al gruppo visto in 2.2.3) è da considerare tipologicamente simile alle lingue suddette; anche in ceco compaiono vocali lunghe in sillabe non accentate. Questo fatto mi sembra interessante in quanto esemplifica la 'trasversalità' dei criteri assunti nella suddivisione in tipi e quindi la possibilità di includere una data lingua in più classi in relazione al criterio di volta in volta adottato.

Passiamo al caso (B). Nelle restanti lingue in cui brevità e lunghezza si oppongono nell'ambito della /vocale/ (2.2.3) vige il vincolo dell'accento lessicale. In franco-provenzale, italo-romanzo settentrionale, friulano, danese, olandese, tedesco settentrionale. le opposizioni di quantità vocalica sono limitate alle sillabe accentate. In tali lingue non compaiono vocali lunghe in sillabe non accentate. Se a ciò si aggiunge che l'accento lessicale è una restrizione che agisce anche in svedese, norvegese, islandese settentrionale, bavarese centrale, lingue caratterizzate da opposizioni di quantità operanti nell'ambito della sequenza /vocale+consonante/ (2.2.1) , si deve concludere che nelle lingue europee odierne la situazione indicata con (B) è senza dubbio prevalente rispetto a quella indicata con (A).

3.1.2 Fonotassi della sillaba e della posizione nella parola

Forma della sillaba e posizione nella parola sono fattori che regolano la distribuzione dei fenomeni temporali in modo da consentire ulteriori specificazioni tipologiche. A seconda delle lingue le opposizioni di quantità possono avere una distribuzione 'libera' o una distribuzione 'vincolata', cioè soggetta a restrizioni che sono di varia entità e che talora si intrecciano. Illustreremo questi concetti generali con un breve esame di tre situazioni che vengono separate per ragioni espositive, ma che hanno indubbie relazioni tra di loro.

In primo luogo guardiamo a ciò che avviene in /sillaba accentata aperta all'interno di parola/ (d'ora in poi: sillaba aperta non finale). Nelle lingue prese in considerazione si osserva che la quantità vocalica si comporta ora in modo 'libero' ora in modo 'vincolato'. Sulla base di tale constatazione distribuzionale è possibile suddividere le lingue in due gruppi: (a) finlandese, antico inglese, ceco, franco-provenzale, italo-romanzo settentrionale, friulano, danese; (b) olandese, tedesco settentrionale.

Molte lingue assegnate al primo tipo, in generale, oppongono brevità e lunghezza vocalica in sillaba aperta non finale senza che questa imponga dei limiti. Esse utilizzano sia vocali brevi sia vocali lunghe. Nelle lingue del secondo tipo invece la sillaba aperta non finale vincola la presenza delle due classi quantitative: sono ammesse esclusivamente le vocali lunghe e vietate le vocali brevi. Possono servire da esempi i seguenti dati:

Frignanese					
it. *fetta*	[ˈfata]	ˈCV-CV	it. *fatta*	[ˈfa:ta]	ˈCV:-CV
Danese					
falde	[ˈfalə]	ˈCV-CV	*male*	[ˈmæ:lə]	ˈCV:-CV
Tedesco settentrionale					
*	*[ˈmɪtə]	*ˈCV-CV	*Miete*	[ˈmi:tə]	ˈCV:-CV
Olandese					
*	*[ˈtɑkən]	*ˈCV-CVC	*taken*	[ˈta:kən]	ˈCV:-CVC

Riguardo a franco-provenzale, italo-romanzo settentrionale e friulano sarà necessario ricorrere alla bibliografia specifica, che descrive e discute le differenziazioni interne alle singole aree, indicando parlate conservative e parlate innovative. Per la situazione friulana e per quella dell'Italo-romània settentrionale disponiamo dei recenti lavori di Miotti (2002, 2007) e di Loporcaro (2005, 2007). Mi limito qui ad un esempio. Nell'italo-romanzo settentrionale l'opposizione tra parole con forma ˈCV-CV e parole con forma ˈCV:-CV è operante in varietà come il cremonese, il frignanese e molte altre. Essa non opera invece in numerose varietà lombarde, come il milanese, dove in sillaba aperta non finale compaiono vocali brevi ma non vocali lunghe: quindi si trovano parole con forma ˈCV-CV, mentre mancano parole con forma ˈCV:-CV, che andrebbe asteriscata.

In secondo luogo, accenniamo alla distribuzione della quantità vocalica in /sillaba accentata chiusa/ (d'ora in poi: sillaba chiusa). A differenza di ciò che si è visto nella situazione descritta sopra, in questa struttura sillabica non si riscontrano vincoli che limitino la distribuzione della brevità e della lunghezza. Nelle lingue esaminate si nota uniformità di comportamento, nel senso che in tutte troviamo opposizioni tra brevi e lunghe quando la sillaba è chiusa: CVC vs. CV:C. Forse non è superfluo menzionare il fatto che l'olandese e il tedesco settentrionale oppongono vocali brevi e vocali lunghe soltanto in sillaba chiusa.

In terzo luogo ci soffermeremo sulla situazione che emerge in /sillaba accentata aperta in fine di parola/ (d'ora in poi: sillaba aperta finale). Forma della sillaba e posizione nella parola si congiungono nel determinare l'organizzazione delle lingue in merito all'uso libero o vincolato delle distinzioni di quantità vocalica. Ne consegue la possibilità di fare una bipartizione tra alcune delle lingue esaminate in questo articolo. Collochiamo da una parte lingue nelle quali la sillaba aperta finale lascia libero l'uso di brevità o lunghezza: il franco-provenzale, l'italo-romanzo settentrionale, il friulano, il danese (in quest'utimo caso troviamo per esempio *nŭ* 'ora'). Da un'altra parte collochiamo lingue in cui la sillaba aperta finale consente esclusivamente l'occorrenza delle vocali lunghe e non quella delle vocali brevi. Questa restrizione agiva già in lingue germaniche antiche come l'antico inglese e l'antico alto tedesco (in questo caso si avevano forme come *sī, nū, jā*); oggi è una caratteristica tipica del tedesco settentrionale. Come esponenti delle due diverse modalità, scelgo il frignanese e il tedesco settentrionale.

Frignanese

it. 'piede'	[pe]	CV	it. 'piedi'	[pe:]	CV:
it. 'su'	[sø]	CV	it. 'suoi'	[sø:]	CV:

Tedesco settentrionale

*	*[fɪ]	*CV	*Vieh*	[fi:]	CV:

4. TERZA NOTA: ASPETTI DELLA PROBLEMATICA DEL TAGLIO SILLABICO

4.1.0 Premessa

Della problematica del taglio sillabico in lingue europee segnaleremo solo qualche aspetto, prendendo le mosse da alcuni fatti significativi visti in parte nei paragrafi precedenti.

A mò di premessa, riprendo e completo le proprietà che contribuiscono a conferire al tedesco settentrionale una fisionomia peculiare. Esse possono essere riassunte in quattro punti: (1) le vocali brevi non possono ricorrere in sillaba aperta all'interno di parola; (2) le vocali brevi non possono ricorrere in sillaba aperta in fine di parola; (3) le vocali brevi si oppongono a vocali lunghe solo in sillaba chiusa; (4) nella sequenza 'vocale tonica breve+consonante breve+vocale atona' la consonante deve essere ambisillabica.

Queste caratteristiche fanno parte di quella che è stata denominata la *Silbenschnitt-syndrom*, nella cornice generale di una teoria che interpreta le opposizioni di quantità vocalica di alcune lingue germaniche come opposizioni prosodiche chiamate ora *Silbenschnittgegensätze* ora *Anschlussartgegensätze.*

4.1.1 I concetti di 'Silbenschnitt' *e di* 'Anschlussart'

Anche se la bibliografia odierna preferisce parlare della teoria del *Silbenschnitt (syllable cut*), per ragioni di chiarezza in questo excursus faremo talvolta riferimento a entrambi i concetti, fra loro connessi, di *Silbenschnitt* (taglio sillabico) e di *Anschlussart* (modo di legame). Essi offrono agli studiosi un rilevante criterio tipologico che viene usato ora in forma forte ora in forma debole. Secondo alcuni il taglio sillabico e il modo di legame autorizzano una netta dicotomia fra due tipi di lingue: le *Quantitätensprachen* e le *Silbenschnittsprachen*; secondo altri essi forniscono un metodo per descrivere adeguatamente modalità di differenziazione temporale che caratterizzano un tipo speciale di sistema quantitativo.

Gli accenni che faremo qui possono essere integrati con gli appunti più particolareggiati di Uguzzoni (2002), che illustrano, con citazioni dirette, coppie concettuali e terminologiche come*: Stark geschnittener Silbenaccent* vs. *Schwach geschnittener Silbenaccent, Fester Anschluss* vs. *Loser Anschluss, Scharfer Schnitt* vs. *Sanfter Schnitt* (Sievers, 1893; Jespersen, 1904; Trubetzkoy, 1939).

La fenomenologia del taglio sillabico si trova soltanto in presenza dell'accento lessicale e consiste nell'opporre alla prosodia chiamata *scharfer Schnitt* (taglio 'brusco') la prosodia chiamata *sanfter Schnitt* (taglio 'piano'). Il membro marcato della opposizione prosodica di taglio sillabico è il taglio brusco, che, come diremo, si combina con la brevità della vocale (Trost, 1939). Nelle lingue che non hanno l'opposizione di taglio sillabico di solito compare soltanto il membro non marcato, cioè il taglio piano. In ceco per esempio sia parole con vocali brevi sia parole con vocali lunghe presentano il taglio piano (Vennemann, 2000).

La realizzazione di coppie di parole tedesche come *Mitte* [ˈmɪtə] vs. *Miete* [ˈmi:tə], *Bett* [bɛt] vs. *Beet* [be:t] coinvolge vari elementi segmentali e intersegmentali: la quantità e la qualità della vocale accentata, la consonante postonica, l'interazione tra vocale accentata e consonante successiva. Nel caso di *scharfer Schnitt* (*abrupt cut*, *coupe ferme*) la vocale è sempre breve e di solito rilassata; la vocale è tagliata dalla consonante successiva ed è coarticolata fortemente con questa*: fester Anschluss* (*close contact*). Invece nel caso di *sanfter Schnitt* (*smooth cut*, *coupe lâche*)) la vocale è sempre lunga e di solito tesa; la vocale completa il suo corso ed è coarticolata debolmente con la consonante che la segue: *loser Anschluss* (*loose contact*). Un'altra differenza fonetica tra parole con taglio brusco e parole con taglio piano è questa: la forza (*Stärke*) e la durata della consonante postonica sono maggiori nel primo caso rispetto al secondo caso (Adelung, 1790; Becker, 1998, 2002).

4.1.2 Vocale breve e taglio sillabico brusco

In alcune versioni della teoria (Trubetzkoy, 1939; Vennemann 1991, 2000) le differenze tra vocali brevi e vocali lunghe vengono considerate proprietà fonetiche concomitanti che partecipano alla opposizione prosodica di taglio sillabico. Come si è visto, si ha una sistematica combinazione di vocale breve, taglio sillabico brusco, legame forte; e si deve riconoscere che la brevità della vocale accentata è assicurata e/o corroborata dal fatto che la vocale viene tagliata dalla consonante successiva e legata strettamente con questa.

I vincoli che agiscono nella distribuzione della brevità vocalica nel tedesco settentrionale assumono una configurazione diversa nel quadro della teoria del taglio sillabico. Alle caratteristiche distribuzionali elencate sopra (4.1.0) fanno da parallelo le seguenti: (1) il taglio brusco non è ammesso in sillaba aperta all'interno di parola; (2) il taglio brusco non è ammesso in sillaba aperta in fine di parola; (3) il taglio brusco è ammesso in sillaba genuinamente chiusa (per 'natura'); (4) il taglio brusco è ammesso in sillaba virtualmente chiusa (per 'ambisillabicità').

Come risulta evidente, queste condizioni hanno un denominatore comune: l'obbligo che la vocale sia seguita nell'ambito della stessa parola da una consonante, cosa che è necessaria per produrre il taglio brusco e il legame forte. Secondo Becker (1998, 2002) l'obbligatorietà della consonante è un fattore centrale per la opposizione di taglio sillabico. È stato messo in rilievo da Vennemann (2000) che nelle lingue germaniche la correlazione naturale e preferita fra struttura sillabica e modi di taglio sillabico in generale si riassume in questa formula: taglio piano in sillabe aperte, taglio brusco in sillabe chiuse. Ma ciò non implica l'impossibilità che nella storia di una lingua si formi taglio brusco in sillabe aperte.

La produzione di taglio brusco in una sillaba aperta in effetti è possibile, ma soltanto a patto che ad essa segua un'altra sillaba nell'ambito della stessa parola. La consonante di questa sillaba fornisce il mezzo indispensabile per interrompere la vocale della sillaba precedente. Tale consonante, che è semplice, viene associata nello stesso tempo sia alla vocale accentata della prima sillaba sia alla vocale non accentata della seconda sillaba. La ambisillabicità può essere rappresentata con qualche espediente grafico: il grafo che userò qui mi sembra iconico e va scritto sotto alla consonante intervocalica (Ç).

Allo scopo di chiarire meglio questo aspetto riprendo gli stessi esempi del paragrafo 3., modificandone la interpretazione alla luce della teoria del taglio sillabico. Ciò riguarda ovviamente la parte sinistra di quel quadro (in 3.1.2), dove sono asteriscate, in quanto escluse dalla fonotassi delle lingue in questione, forme bisillabiche con vocale breve in sillaba aperta all'interno di parola *ˈCV-CV(C). Secondo la prospettiva indicata ora. le parole, rispettivamente del tedesco settentrionale e dell'olandese, *Mitte* e *takken* vanno analizzate come forme con vocale interrotta dalla consonante successiva, la quale conferisce alla sillaba accentata una sorta di chiusura 'virtuale'. In questa maniera [ˈmɪt̪ə] e [ˈtɑk̪ən] risultano conformi alla regola fonotattica della *Silbenschnittsyndrom*: ˈCVÇV(C), vista nel punto (4) del presente paragrafo.

Tedesco settentrionale					
Mitte	[ˈmɪt̪ə]	ˈCVÇV	*Miete*	[ˈmi:tə]	ˈCV:-CV
Olandese					
takken	[ˈtɑk̪ən]	ˈCVÇVC	*taken*	[ˈta:kən]	ˈCV:-CVC

Non intendo però trascurare l'opinione di Martinet (1966, 1975) sul taglio sillabico inteso come il modo in cui si comporta dal punto di vista della divisione delle sillabe una consonante collocata fra due vocali. Secondo Martinet in primo luogo il fenomeno riguarda essenzialmente bisillabi come tedesco *Kämmen*, più che monosillabi come tedesco *Kamm*; in secondo luogo il taglio brusco (*coupe ferme*) si ha quando la consonante interna appartiene almeno tanto alla prima sillaba quanto alla seconda. La [m] di *Kämmen* aderisce alla vocale breve precedente e rende la prima sillaba della parola "une syllabe incontestabilment fermée ou entravée" e la trascrizione proposta è [ˈkɛm-ən] (Martinet, 1975: 187)

4.1.3 Quantità vocalica e taglio sillabico

Per procedere a qualche altra considerazione sul tema del taglio sillabico mi sembra opportuno premettere un brano di Martinet (1975: 189): "l'existence de la coupe ferme en anglais, en néerlandais et en allemand contemporaines n'implique pas que ce phénomène existait déjà lorsque ces trois langues n'en faisaient qu'une, mais que cette langue commune comportait les traits qui, par action des uns sur les autres et dans le cadre général des besoins communicatifs de l'humanité, devaient faire apparaître la coupe ferme".

Vediamo ora brevemente quando e come può essersi formata l'opposizione di taglio sillabico che oggi caratterizza il tedesco settentrionale. Secondo l'ipotesi più condivisa questa innovazione si è verificata nel passaggio dal medio alto tedesco al nuovo alto tedesco, quando, per varie ragioni, le opposizioni di quantità vocalica si trovarono in pericolo di essere soppresse. Il ruolo principale viene attribuito all'azione di un forte accento dinamico o espiratorio. Al fine di mantenere una distinzione tra vocali brevi e vocali lunghe in sillabe accentate, si sentì la necessità di controbattere l'effetto allungante di tale accento. La brevità della vocale accentata venne 'salvata' grazie a quei processi di interazione tra vocale e consonante di cui si è parlato: il taglio brusco e il legame forte.

Anche a questo proposito ritengo utile una citazione testuale che sintetizza la questione: "das Deutsche ist zu der Zeit zur Silbenschnittsprache geworden, als seine Phonotaktik (das Verbot kurzer offener Tonsilben) es den Sprechern erlaubte, Vokalkürze durch Koartikulation mit den Folgekonsonanten herzustellen" (Becker, 1998: 72).

L'accenno fatto all'accento dinamico porta con sé almeno due osservazioni. Come sappiamo, Sievers (1893) introduce la sua distinzione fra *stark geschnittener Silbenaccent e schwach geschnittener Silbenaccent* nell'ambito della trattazione dello *expiratorischer oder dynamischer Silbenaccent*; egli sottolinea che la prima modalità (corrispondente ai concetti più noti di *scharfer Schnitt* e di *fester Anschluss*) si trova solo in lingue che hanno un forte accento espiratorio o dinamico. D'altra parte ricordiamo che ricerche recenti su inglese, olandese, tedesco hanno 'riabilitato' posizioni tradizionali che classificavano come dinamico l'accento germanico e vedevano il suo correlato primario in un grado elevato di intensità percepita; e si è aperta la strada per in nuova luce concetti come quello di *expiratorische Verstärkung* che per tradizione era considerato un correlato fisiologico dell'accento in alcune lingue germaniche (Uguzzoni 2006b: 106, 110).

Bisogna però a questo punto fare una precisazione. Se si allarga lo sguardo a lingue non germaniche e non europee si osserva che l'opposizione di taglio sillabico non è necessariamente collegata né con fatti quantitativi né con la prominenza accentuale. Pertanto non si può generalizzare l'ipotesi che le distinzioni di taglio sillabico siano il risultato di una riorganizzazione di distinzioni di quantità vocalica dovuta a difficoltà che l'accento lessicale e in particolare l'accento dinamico può creare.

Infine si deve tenere presente che l'opposizione di taglio sillabico è limitata ad alcune varietà del tedesco contemporaneo. La distribuzione areale del fenomeno è trattata, per esempio, nel libro di Spiekermann (2000).

4.1.4 Il problema dei correlati fonetici

Per molto tempo, a cominciare da Valentin Ickelsamer, 1534 (citato da Becker, 1998: 157), i due diversi modi di interazione tra vocale tonica e consonante postonica sono stati descritti in termini intuitivi e impressionistici (Becker, 1998; Mooshammer, 1998; Restle, 2002). Informazioni sulle principali ricerche sperimentali svolte dal 1962 al 2003 sono contenute in Uguzzoni (2002, 2006a, 2006b) e in Uguzzoni *et al.* (2003). In questa sede mi limito a sintetizzare alcuni lavori condotti esplicitamente con lo scopo di cercare nella sostanza fisiologica e/o acustica correlati fonetici della impressione uditiva di legame forte vs. debole e di taglio brusco vs. piano.

Sulle differenze fonetiche tra *fester* e *loser Anschluss* (*close* e *loose contact*) nel tedesco settentrionale è stata compiuta una indagine approfondita da Fischer-Jørgensen & Jørgensen (1969), fondata su un corpus di 2066 parole piane e tronche terminanti in consonante. A conclusione della loro analisi sperimentale su sei soggetti maschili, che davano l'impressione di produrre tutti, anche se con gradi diversi il legame forte tra vocale breve accentata e consonante successiva, gli autori hanno rilevato che non era stato possibile trovare né a livello fisiologico né a livello acustico una dimensione specifica e indipendente valevole come sostrato oggettivo del fenomeno percettivo in esame. Si è pertanto avanzata l'ipotesi che alla base della distinzione fra legame forte e legame debole ci sia un plausibile intreccio di parametri fonetici già noti.

In anni più recenti Spiekermann (2000, 2002) è andato alla ricerca di uno specifico correlato acustico della opposizione di taglio sillabico nel tedesco settentrionale. Come abbiamo visto nelle pagine precedenti, non si è mai dubitato del fatto che il taglio brusco

(*scharfer Schnitt*) e il taglio piano (*sanfter Schnitt*) sono connessi, rispettivamente, con la brevità e con la lunghezza della vocale. Ma siccome la durata è il correlato fonetico anche della opposizione di quantità vocalica, l'autore ha ritenuto indispensabile cercare di individuare un altro parametro che consenta di differenziare a sul piano acustico le *Silbenschnittsprachen* dalle *Quantitätensprachen.*

La sua indagine verte sulle caratteristiche del profilo dell'energia nella vocale accentata di parole trocaiche con una singola consonante intervocalica. Vengono analizzate tre proprietà: il numero dei picchi dell'energia, la posizione del picco dell'energia quando la vocale ha un solo picco, la forma del tracciato dell'energia prima e dopo un picco. Dal confronto fatto da Spiekermann con dati del finlandese e del ceco emerge che la prima proprietà si riscontra anche in queste lingue, dove le vocali lunghe hanno un numero di picchi maggiore rispetto alle vocali brevi. Risulta invece che le differenze associate con la seconda e con la terza proprietà si trovano soltanto in tedesco. Si conclude che la posizione di un unico picco di energia e la forma del tracciato dell'energia costituiscono i correlati specifici e rilevanti della opposizione di taglio sillabico operante nel tedesco settentrionale.

Mentre lo studio di Spiekermann si occupa dell'andamento dell'energia lungo l'asse del tempo e si fonda su misurazioni dell'energia globale (*overall signal amplitude*), Jessen (2002) ha impostato il suo lavoro su un nuovo modo di misurare l'energia *(frequency - sensitive amplitude measurements*). L'analisi di parole tedesche in cui in posizione accentata compaiono /ɪ, ɛ, ʊ, ɔ, a/ (*abruptly cut vowels*) e /i:, e:, u:, o:, a:/ (*smoothly cut vowels*) ha permesso a Jessen di affermare che tali vocali (a eccezione delle basse) si distinguono tra di loro in maniera statisticamente significativa per valori differenti dei parametri H1-A2 e H1-A3: (per i particolari tecnici rinvio a Uguzzoni, 2006a e 2006b). I valori della sottrazione risultano minori nelle vocali con taglio brusco rispetto a ciò che si trova nelle corrispondenti vocali con taglio piano. Valori più piccoli di H1-A2 e H1-A3 indicano che una classe di vocali presenta nel campo delle medie e alte frequenze un grado di energia maggiore rispetto alla classe di vocali contrapposta. Il *new look* , che è cominciato a metà degli anni '90 e a cui si può dare il nome 'intensità rivisitata' (Uguzzoni, 2003, 2006a, 2006b), si è rivelato importante anche nel caso dello studio della opposizione di taglio sillabico. Fra taglio brusco e taglio piano ci sono differenze di intensità caratterizzabili in termini di *spectral balance* o di *mid-to-high-frequency emphasis*. In sintonia con questa conclusione si può ipotizzare che le vocali della prima categoria siano percettivamente più forti (*loud*) di quelle della seconda categoria (Uguzzoni, 2006; Uguzzoni, 2006b: 114-117).

Sul piano articolatorio sono state condotte interessanti ricerche di carattere cinematico grazie a recenti strumentazioni come ELITE e EMMA. In alcuni casi le risultanze cinematiche delle indagini svolte in Germania sulle vocali brevi e rilassate vs. le vocali lunghe e tese del tedesco standard sono state interpretate come valide anche per l'opposizione di taglio sillabico (Hertrich & Ackermann, 1997; Kroos *et al.*, 1997; Hoole & Mooshammer, 2002).

Zmarich *et al.* (2003), Zmarich & Uguzzoni (2006) hanno fatto analisi cinematiche dei gesti labiali in tre tipi di parole: bisillabi con l'accento nella prima sillaba, /ˈpV(:)pa/; monosillabi terminanti in consonante, /pV(:)p/; monosillabi terminanti in vocale, /pV(:)/ (in questo caso la consonante iniziale della parte destra della frase cornice era /p/). Parole e pseudo-parole fonotatticamente ammissibili in frignanese sono state prodotte da un parlante nativo da tre a cinque volte. Le misurazioni di vari aspetti della cinematica dei gesti labiali

hanno portato a individuare una serie di differenze sistematiche. Ciò ha consentito di gettare luce sul comportamento articolatorio associato con vocali brevi (/e, ø, ɔ, a/) vs. il comportamento articolatorio associato con vocali lunghe (/e:, ø:, ɔ:, a:/). Per i risultati specifici dell'esame di alcuni parametri si rinvia, oltre che a Zmarich *et al.* (2003) , a Uguzzoni *et al.*, 2003 e Uguzzoni, 2006b (:121-124), dove si mettono in rilievo fenomeni di coordinazione temporale di gesti articolatori adiacenti (*overlap* e *truncation*).

4.1.5 Distinzione di taglio sillabico anche nell'italo-romanzo settentrionale?

L'accenno a questi lavori cinematici riguardanti un dialetto italo-romanzo di area emiliana può sembrare fuori posto nell'ambito di una nota dedicata al taglio sillabico. Lo sarebbe davvero se effettivamente il frignanese avesse una opposizione di quantità vocalica di tipo 'classico', non confrontabile con l'opposizione di taglio sillabico del tedesco settentrionale. Ma, come parlante frignanese, ho l'impressione che le cose non stiano proprio così (Uguzzoni *et al.*, 2003; Uguzzoni, 2006b: 117-121).

Sono del parere che varrebbe la pena continuare studi e ricerche per chiarire se, per fare un esempio, [ˈfata] del frignanese ha più 'somiglianze' con [ˈfalə] (CV-CV) del danese o con [ˈmɪt̪ə] (ˈCVÇV) del tedesco. Ho fatto riferimento appositamente a bisillabi piani perché la vera 'quaestio' riguarda appunto questo tipo di parole. Nella prima ipotesi [ˈfata] avrebbe la struttura ˈCV-CV (3.1.2), nella seconda ipotesi invece dovrebbe avere la struttura ˈCVÇV (4.1.2).

Un fatto frignanese che in me e in altri ha sempre suscitato una certa sorpresa riguarda la misura della durata della consonante successiva alla vocale accentata all'interno di parole bisillabiche ˈCV(:)CV. Come ho sottolineato in 3.1.2, l'estensione temporale della consonante postonica non manifesta differenze degne di nota in parole come [ˈmela], [ˈtøla], [ˈtɔka], [ˈsapa], dove la vocale è breve, da un lato, e in parole come [ˈme:la], [ˈtø:la], [ˈtɔ:ka], [ˈsa:pa], dove la vocale è lunga, dall'altro. Il problema consiste nella discrasia tra il risultato strumentale e l'impressione uditiva. Infatti si ha la sensazione che la consonante successiva a vocale breve sia un po' più lunga e un po' più forte rispetto alla consonante successiva a vocale lunga.

È lecito chiedersi: l'impressione uditiva di una maggiore salienza della consonante nella condizione /ˈCVCV/ può essere messa in relazione con la sindrome del taglio sillabico e del modo di legame? In parole come [ˈmela], [ˈtøla], [ˈtɔka], [ˈsapa] sembra che alla brevità della vocale accentata si accompagni "qualche altra cosa". Nella condizione /ˈCVCV/ l'orecchio di nativi e non nativi 'sente' che l'interazione tra vocale accentata breve e consonante successiva è dello stesso genere illustrato nelle pagine precedenti: un legame forte e un taglio brusco. Un'altra domanda: si può ipotizzare che i meccanismi intergesturali (*overlap* e *truncation*) emersi dall'analisi cinematica di sequenze frignanesi /p+V+p/ siano soggiacenti non solo alla brevità vocalica misurata acusticamente, ma anche alla sensazione di legame forte e di taglio busco?

RINGRAZIAMENTI

Esprimo la mia viva gratitudine a Stephan Schmid, Renzo Miotti, Antonella Giannini e ai miei compaesani: in particolare gli abitanti delle frazioni di Pavullo nel Frignano, a cui desidero dedicare queste note zurighesi.

5. BIBLIOGRAFIA

Adelung, J. Ch. (1790), *Vollständige Anweisung zur deutschen Orthographie nebst einem kleinen Wörterbüche für die Aussprache, Orthographie, Biegung und Ableitung*, Leipzig: Weigand.

Allen, W. S (1973), *Accent and rhythm. Prosodic features of Latin and Greek. A study in theory and reconstruction*, Cambridge: Cambridge University Press.

Árnason, K. (1980), *Quantity in historical linguistics. Icelandic and related cases*, Cambridge: Cambridge University Press.

Auer, P., Gilles. P. & Spiekermann, H. (2002), Introduction: Syllable cut an tonal accents. Two 'exceptional prosodies' of Germanic and some thoughts on their mutual relationship. in *Silbenschnitt und Tonakzente* (P. Auer *et al*. editors), Tübingen: Niemeyer, 1-10.

Bannert, R. (1976*), Mittelbairische Phonologie auf akustischer and perzeptorischer Grundlage*, Lund: Gleerup/München: Fink.

Bannert, R. (1977), Quantität im Mittelbairischen: Komplementäre Länge von Vokal und Konsonant, in *Phonologica* 1976 (W.U. Dressler & O.E. Pfeiffer, editors), Innsbruck: Institut für Sprachwissenschaft, 261-270.

Baroni, M. & Vanelli, L. (2000), The relationship between vowel length and consonantal voicing in Friulian, in *Phonological theory and the Dialects of Italy* (L. Repetti, editor), Amsterdam: Benjamins, 13-44.

Becker, T. (1998), *Das Vokalsystem der deutschen Standardsprache*, Frankfurt: Peter Lang.

Becker, T. (2002), Silbenschnitt und Silbenstruktur in der deutschen Standarsprache der Gegenwart, in *Silbenschnitt und Tonakzente* (P. Auer *et al.*, editors), Tübingen: Niemeyer, 87-101.

Bertinetto, P.M. (1981), *Strutture prosodiche dell'italiano*, Firenze: Accademia della Crusca.

Bosoni, G. (1995), Dialettologia lombarda: un esempio di approccio strumentale allo studio delle opposizioni di quantità vocalica in sillaba tonica, *Studi italiani di Linguistica Teorica e Applicata*, 24, 345-364.

Coco, F. (1970), *Il dialetto di Bologna. Fonetica storica e analisi strutturale*, Bologna: Forni.

Elert, C.-Ch. (1964), *Phonologic studies of quantity in Swedish*, Uppsala: Almqvist & Wiksell.

Eliasson, S. (1978), Swedish quantity revisited, in *Nordic prosody. Papers from a symposium* (E. Gårding et al., editors), Travaux de l'Institut de Linguistique de Lund 13, Lund: Lund University, 111.122.

Fallows, D. (1981), Experimental evidence for English syllabification and syllable structure, *Journal of Linguistics*, 17, 309-317.

Fischer-Jørgensen, E. (1972), Formant frequencies of long and short Danish vowels, in *Studies for Einar Haugen* (E. Firchow *et al.*, editors), The Hague: Mouton.

Fischer-Jørgensen, E. (1990), Intrinsic F0 in tense and lax vowels with special reference to German, *Phonetica*, 47, 99-140.

Fischer-Jørgensen, E. & Jørgensen, H. P. (1969), Close and loose contact (Anschluss) with special reference to North German, *Annual Report of the Institute of Phonetics of the University of Copenhagen*, 4, 43-80.

Francescato, G. (1960), *Dialettologia friulana*, Udine/Tolmezzo: Società Filologica Friulana.

Hajek, J. (1997), Analisi acustica delle quantità segmentali in area Bolognese, Rivista Italiana di Dialettologia, 21, 133-147.

Hakulinen, L. (1957), *Handbuch der finnischen Sprache*, Wiesbaden: Harrassowitz.

Harrington, J., Flechter, J. & Roberts, C. (1995), Coarticulation and the accented-unaccented distinction: evidence from jaw movement, *Journal of Phonetics*, 23, 305-322.

Herman, J. (1990), *Du latin aux langues romanes. Etudes de linguistique historique*, Tübingen: Niemeyer.

Hertrich, I. &Ackermann, H. (1997), Articulatory control of phonological vowel length contrasts. Kinematic analysis of labial gestures, *Journal of the Acoustical Society of America*, 102, 523-536.

Hoole, P. & Mooshammer, Ch. (2002), Articulatory analysis of the German vowel system, in *Silbenschnitt und Tonakzente* (P. Auer et al., editors), Tübingen: Niemeyer, 129-152.

Hurch, B., &Tonelli, L. (1982), /'matto/ oder /'mat:o/? Jedenfalls ['mat:o], Zur Konsonantenlänge im Italienischen, *Wiener Linguistische Gazette*, 29, 17-38.

Jessen, M. (2002), Spectral balance and its relevance for syllable cut theory, in *Silbenschnitt und Tonakzente* (P. Auer *et al*., editors), Tübingen: Niemeyer, 153-179.

Kroos, C., Hoole, Ph., Kühnert, B. & Tillmann, H.G. (1997), Phonetic evidence for the phonological status of the tense-lax distinction in German, *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, 35, 17-25.

Kučera, H. (1961), *The phonology of Czech*,'s-Gravenhage: Mouton.

Ladefoged, P. (2003), *Phonetic data analysis. An introduction to fieldwork and instrumental techniques*, Oxford: Blackwell.

Lehiste, I. (1970), *Suprasegmentals*, Cambridge, MA: MIT Press.

Lehtonen, J. (1970), *Aspects of quantity in Standard Finnish*, Jyväskylä: Jyväskylä University Press.

Linell, P. (1978), Vowel length and consonant length in Swedish word phonology, in *Nordic prosody. Papers from a symposium* (E. Gårding et al., editors), Travaux de l'Institut de Linguistique de Lund 13, Lund: Lund University, 123-136.

Loporcaro, M. (2005), La lunghezza vocalica nell'Italia settentrionale alla luce dei dati del lombardo alpino, in *Atti del convegno di dialettologia in onore del prof. Remo Bracchi*, Bormio, 24-25 settembre 2004 (M. Pfister & G. Antonioli, editors), Istituto di Dialettologia e di Etnografia Valtellinese e Valchiavennasca, 97-113.

Loporcaro, M. (2007), Facts, theory and dogmas in historical linguistics. Vowel quantity from Latin to Romance, in *Historical Linguistics 2005*, Amsterdam: John Benjamins, 311-336.

Loporcaro, M., Delucchi, R., Nocchi, N., Paciaroni, T. & Schmid, S. (2006), La durata consonantica nel dialetto di Lizzano in Belvedere (Bologna), in *Analisi prosodica. Teorie, modelli e sistemi di annotazione* (R. Savy & C. Crocco, editors), Atti del 2° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Salerno, 30 novembre–2 dicembre 2005, Torriana (RN): EDK Editore, 491-517 (CD-ROM).

Loporcaro, M., Paciaroni, T. & Schmid, S. (2005), Consonanti geminate in un dialetto lombardo alpino, in *Misura dei parametri. Aspetti tecnologici ed implicazioni nei modelli linguistici* (P. Cosi, editor), Atti del 1° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Padova, 2-4 dicembre 2004, Torriana (RN): EDK Editore, 597-618.

Lüdtke, H. (1956), *Die strukturelle Entwicklung des romanischen Vokalismus*, Diss. Bonn: Romanisches Seminar der Universität Bonn.

Malmberg, B. (1944), *Die Quantität als phonetisch-phonologischer Begriff. Eine allgemein-sprachliche Stud*ie, Lund: Gleerup.

Malmberg, B. (1949), *Voyelles longues et voyelles brèves*, Studia Linguistica, 9, 80-87.

Martinet, A. (1956), *La description phonologique, avec application au parler franco-provençal d'Hauteville (Savoie),* Genève: Droz/Paris: Minard.

Martinet, A. (1966), Close contact, *Word*, 22, 1-6.

Martinet, A. (1969), Coupe ferme et coupe lâche, in *Mélanges pour Jean Fourquet* (P. Valentin & G. Zinke, editors), Paris: Klincksieck, 221-226.

Martinet, A. (1975), La coupe ferme en germanique, in A. Martinet, *Evolution des langues et reconstruction*, Paris: PUF,185-193.

Martinet, A. (1988), Où commencent les structures syllabiques de Haute-Maurienne? in *Espaces romans. Etudes de dialectologie et de géolinguitique offerts à Gaston Tuaillon*, Grenoble: Ellug, 1, 158-163.

Mayerthaler, E. (1996), Stress, syllables, and segments: their interplay in Italian dialect continuum, in *Natural Phonology. The State of the Art* (B. Hurch & A Rhodes, editors), Berlin: Mouton de Gruyter, 201-221.

Miotti, R. (2007), Le varietà di Dignano, Flaibano e Sedegliano nel contesto dei dialetti friulani. Aspetti fonologici, in *Ladine loqui. IV Colloquium Retoromanistich*, San Denêl ai 26 e 27 di avost dal 2005 (F. Vicario, editor), Udine: Società Filologica Friulana, 71-117.

Miotti. R. (2002), Friulian, *Journal of the International Phonetic Association*, 32, 237-247.

Mooshammer, Ch. (1998), Experimentalphonetische Untersuchungen zur artikulatorischen Modellierung der Gespanntheitopposition im Deutschen, *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, 36, 3-192.

Muljačić, Ž. (1972), *Fonologia della lingua italiana*, Bologna: Il Mulino.

Murray, R.W. (2000), Syllable cut prosody in Early Middle English, *Language*, 76, 617-654.

Ramers, K. (1988), *Vokalquantität und -qualität im Deutschen*, Tübingen: Niemeyer.

Repetti, L. (1992), Vowel length in Northern Italian dialects, Probus, 4, 155-182.

Restle, D. (2002), Normierung der Silbenquantität. Ein typologischer Beitrag zur Charakteristik des Silbenschnitts in und ausserhalb der Germania, in *Silbenschnitt und Tonakzente* (P. Auer *et al*., editors), Tübingen: Niemeyer, 35-66.

Riad, T. (1995), The quantity shift in Germanic: a typology, in H. Fix (editor*), Quantitätsproblematik und Metrik*, Amsterdamer Beiträge zur älteren Germanistik, 42, Amsterdam: Rodopi, 160-184.

Sanga, G. (1988), La lunghezza vocalica nel milanese e la coscienza fonologica dei parlanti, *Romance Philology*, 41, 290-297.

Schmid, S. (1997), A typological view of syllable structure in some Italian dialects, in *Certamen Phonologicum* III (P.M. Bertinetto *et al*., editors), Torino: Rosenberg-Sellier, 247-265.

Sievers, E. (1893), *Grundzüge der Phonetik zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen*, Leipzig: Breitkopf & Härtel.

Spiekermann, H. (2000), *Silbenschnitt in Deutschen Dialekten*, Tübingen: Niemeyer.

Spiekermann, H. (2002), Ein akustisches Korrelat des Silbenschnitt: Formen des Intensitätsverlaufs in Silbenschnitt- und Tonakzentsprachen, in *Silbenschnitt und Tonakzente* (P. Auer *et al*., editors), Tübingen: Niemeyer, 181-199.

Suomi, K. (2005), Temporal conspiracies for a tonal end. Segmental durations and accentual F0 movement in a quantity language, *Journal of Phonetics*, 33, 291-309.

Trost, P. (1939), *Bemerkungen zum deutschen Vokalsystem*, Travaux du Cercle Linguistique de Prague, 8, 319-326.

Trubetzkoy, N. S. (1939), *Grundzüge der Phonologie*, Travaux du Cercle Linguistique de Prague, 7, 1-268.

Trubetzkoy, N.S. (1938), Die phonologischen Grundlagen der sogenannten 'Quantität' in den verschiedenen Sprachen, in *Scritti in onore di Alfredo Trombetti*, Milano: Hoepli, 155-174.

Uguzzoni, A. (1971), Quantità fonetica e quantità fonematica nell'area dialettale frignanese, *L'Italia Dialettale*, 34, 115-136.

Uguzzoni, A. (1974), Sulla struttura della parola dei dialetti emiliani: aspetti sincronici e aspetti diacronici di un problema, *Atti e Memorie della Deputazione di Storia Patria per le Antiche Provincie Modenesi*, 9, 239-252.

Uguzzoni, A. (1975), Appunti sulla evoluzione del sistema vocalico di un dialetto frignanese, *L'Italia Dialettale*, 38, 47-76.

Uguzzoni, A. (1979), Sulla struttura della parola dei dialetti emiliani: aspetti sincronici e aspetti diacronici di un problema, in *Atti del XIV Congresso Internazionale di Linguistica e Filologia Romanza*, Napoli, 15-20 aprile, 1974, Napoli: Macchiaroli/Benjamins, 3, 105-115.

Uguzzoni, A. (1990), Long and short vowels in Frignano dialects, The role of past and present syntagmatic dimension, *Studi italiani di linguistica teorica e applicata*, 19, 533-547.

Uguzzoni, A. (1997), Phénomènes de durée dans certains parlers de l'Italie du Nord, in *Proceedings of the 16th International Congress of Linguists*, Oxford: Pergamon-Elsevier, CD-ROM, Paper No. 0190.

Uguzzoni, A. (2000a), Aspetti -emici e aspetti -etici del fenomeno della 'quantità vocalica' nei dialetti dell'Italia settentrionale, in *Atti delle X Giornate di Studio del Gruppo di Fonetica Sperimentale*, Napoli: Istituto Universitario Orientale, 227-235.

Uguzzoni, A. (2000b), Fonologia e fonetica della quantità vocalica in area italo-romanza, Il caso dei dialetti del Medio Frignano (provincia di Modena), *Studi Orientali e Linguistici*, 7, 339-349.

Uguzzoni, A. (2002), Fester vs.. loser Anschluss, Appunti per una storia di un concetto secolare, Lingue e Linguaggio, 1, 327-340.

Uguzzoni, A. (2003), In margine ad una rivisitazione della intensità, in *Voce, canto, parlato. Studi in onore di Franco Ferrero* (P. Cosi *et al*., editors), Padova: Unipress, 299-302.

Uguzzoni, A. (2006a), I valori di H1-A2 e H1-A3 come correlati della intensità 'rivisitata'. Aspetti e problemi, in *Analisi prosodica. Teorie, modelli e sistemi di annotazione*, Atti del 2° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Salerno, 30 novembre – 2 dicembre 2005 (R. Savy & C. Crocco, editors), Torriana (RN): EDK Editore., 566-592 (CD-ROM).

Uguzzoni, A. (2006b), Produzione, acustica, percezione della 'intensità rivisitata'. Ricerche in area germanica e in area italo-romanza, *Rivista Italiana di Dialettologia*, 30, 103-138.

Uguzzoni, A. (2009), Vocali accentate brevi prodotte con incremento di F0: come e perché?, in *La Fonetica Sperimentale. Metodo e Applicazioni* (L. Romito, V. Galatà & R. Lio, editors), Atti del 4° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Università della Calabria, 3-5 dicembre, 2007, 149-164.

Uguzzoni, A. & Busà, M.G. (1995a), Acoustic correlates of vowel quantity contrasts in an Italian dialect, in *Proceedings of the XIII International Congress of Phonetic Sciences*, Stockholm, Sweden, August 13-19, 1995, 390-393.

Uguzzoni, A. & Busà, M.G. (1995b), Correlati acustici della opposizione di quantità vocalica in area emiliana, *Rivista Italiana di Dialettologia*, 19, 7-39.

Uguzzoni, A., Pettorino, M. & Filipponio, L. (1999), On stressed vowel durations, vowel-consonant contact types and syllable shapes in the Italo-Romance area, in *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, USA, August 1-7, 1999, 2209-2210.

Uguzzoni, A., Azzaro, G. & Schmid, S. (2003), Short vs. long and/or abruptly vs. smoothly cut vowels. New perspectives on a debated question, in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2717-2220.

Vennemann, T. (1991), Syllable structure and syllable cut prosodies in modern standard German, in *Certamen Phonologicum II*. Papers from the 1990 Cortona Phonology Meeting (P.M. Bertinetto *et al*., editors), Torino: Rosenberg & Sellier, 211-243.

Vennemann, T. (2000), From quantity to syllable cuts. On so-called lengthening in the Germanic languages, *Rivista di Linguistica*, 12, 251-282.

Weinrich, H. (1958), *Phonologische Studien zur Romanischen Sprachgeschichte*, Münster: Aschendorffsche Verlagsbuchhandlung.

Zmarich, C., Uguzzoni, A. & Ferrari, V. (2003), Controllo articolatorio della opposizione di quantità vocalica in area emiliana: analisi cinematica dei gesti labiali, in *La coarticolazione* (G. Marotta & N. Nocchi, editors), Atti delle XIII Giornate di Studio del Gruppo di Fonetica Sperimentale, Pisa: Edizioni ETS, 295-306.

Zmarich, C. & Uguzzoni, A. (2006), Confini di sillaba, confini di parola e lunghezza fonologica in area frignanese: Analisi cinematica dei gesti labial, in *Analisi prosodica. Teorie, modelli e sistemi di annotazione* (R. Savy & C. Crocco, editors), Atti del 2° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Salerno, 30 novembre – 2 dicembre 2005, Torriana (RN): EDK Editore, 612-631 (CD-ROM).

FENOMENI D'ARMONIA VOCALICA IN AREA FRIULANA E IBERICA E LE SORTI DI -A FINALE LATINA

Renzo Miotti
Dipartimento di Romanistica - Università di Verona, Italia
renzo.miotti@univr.it

1. SOMMARIO

L'armonia vocalica è un fenomeno assimilatorio che consiste nell'estensione di tutti o d'alcuni tratti d'una vocale ad altri segmenti vocalici, normalmente adiacenti. L'armonia può manifestarsi in una duplice direzione: da una posizione forte, vale a dire prominente dal punto di vista percettivo, verso posizioni più deboli; viceversa, da posizioni poco prominenti verso gli altri segmenti vocalici. In letteratura, le cause dell'armonia vengono riportate al conseguimento di benefici d'ordine strutturale: semplificazione articolatoria (Pulleyblank, 2002), benefici percettivi (soprattutto se da posizioni deboli; cfr. Walker, 2005, 2006), semplificazione articolatoria e benefici percettivi (Cole & Kisseberth, 1994). Per una sintesi della questione, cfr. Jiménez & Lloret (c.d.s.).

Sulla scorta di queste premesse, il presente lavoro intende presentare e discutere i risultati (ancora provvisori, in vista d'essere corroborati da una base più ampia di dati) d'un'indagine condotta sul friulano centro-orientale, che proverebbero l'esistenza di processi riconducibili al primo dei due modelli d'assimilazione (da posizioni forti a posizioni deboli): il tratto [±ATR] (ma non il punto d'articolazione) delle vocali medie semichiuse accentate verrebbe esteso alla vocale media non-accentata finale -/e/ (normalmente [ɛ]) che, diacronicamente, rappresenta il normale esito di -A latina, avvenuto per innalzamento e anteriorizzazione, nella maggior parte dei dialetti centro-orientali, più innovativi (laddove le varietà più marginali conservano -/a/). Va detto che il fenomeno cui si fa riferimento non ha sinora trovato riscontro in letteratura, la quale si limita a rilevare la generale tendenza all'abbassamento ([ẹ, ọ]) delle medie non-accentate in posizione finale (Canepari, 2006[3]; Miotti, 2002), senza però prestar attenzione alle significative differenze riscontrabili, con regolarità, in dipendenza dai condizionamenti visti.

Avremo dunque [ẹ] dopo /ˈe, ˈo/ ma [ɛ] nei contesti non-armonici (cioè dopo qualsiasi altra vocale): [ˈkwẹtẹ, ˈmọte] 'cotta, mossa' ma [ˈsɛ̨dɛ, ˈkɔ̨dɛ] 'seta, coda'. Si tenga presente che pure /ˈe, ˈo/ arrivano a [ˈẹ, ˈọ] (seppure in minor grado rispetto alle non-accentate finali). S'ipotizza che gli stessi meccanismi che operano in sincronia siano stati responsabili, in diacronia, del mutamento lat. -A > friul. centro-orientale moderno -/e/, attraverso i suddetti processi d'innalzamento e anteriorizzazione, a partire da una situazione d'instabilità ed estrema variabilità fonetica che dovette interessare /a/ finale nell'udinese medievale. Il quadro, per queste prime fasi evolutive, sarebbe del tutto simile a quello che è possibile delineare per dialetti valenziani meridionali attuali (vedi oltre).

In friulano centro-orientale, il fenomeno armonico sembrerebbe condizionato da fattori contestuali e da fattori legati alla posizione dei segmenti all'interno dell'enunciato: solo le vocali finali in tonìa (cioè alla fine d'enunciato intonativo, in posizione prepausale) sono interessate dal processo descritto, laddove, all'interno dell'enunciato stesso, l'effetto tende a ridursi fino a scomparire (con una notevole tendenza alla riduzione/centralizzazione dei timbri vocalici, fenomeno che verrà discusso in uno specifico paragrafo).

I fenomeni friulani vengono confrontati con quelli descritti per altre varietà romanze, in particolare iberiche (varietà valenziane meridionali: Jiménez, 1998 e 2001; Jiménez & Lloret, c.d.s.). In valenziano (dove -A latina > -/a/), il processo è limitato a -/a/ preceduta da /ˈɛ, ˈɔ/, che agiscono sulla vocale finale innalzandola e propagando il tratto 'punto d'articolazione' (coronale e labiale, rispettivamente): [isˈtɛɾjɛ] "isteria", [isˈtɔɾjɔ] "storia".

Il confronto friulano-valenziano mette in evidenza alcuni punti problematici d'ordine interpretativo, per quanto riguarda l'interpretazione articolatoria delle cause dei processi esaminati: la semplificazione articolatoria si manifesta all'interno di domini omogenei, con estensione dei tratti a vocali contigue, solo in valenziano, mentre ciò non si verifica necessariamente per il friulano, che conosce minori restrizioni in tal senso. Mentre in valenziano l'armonia interessa solo vocali contigue (nei proparossitoni la vocale postaccentata interna blocca il processo: [ˈtɛtɾika], non *-[kɛ] "tetra"), in friulano, al contrario, non sembrano esserci restrizioni in tal senso. Se in valenziano il dominio dell'armonia è la parola prosodica e le vocali dei pronomi enclitici non sono perciò interessate dal processo armonico, in friulano, al contrario, queste ultime partecipano al processo a livello postlessicale: valenziano [ˈpɛlɛla] (e non *-[lɛ]) "pelala" *vs.* friulano [ˈmẹtilẹ] (e non *-[lɛ]) "mettila".

2. INTRODUZIONE

È noto che -A finale latina (protoromanza) si riflette in una molteplicità di esiti nella Romània (nord-)occidentale[1] moderna, che vanno dalla sua conservazione come vocale bassa [a] fino a realizzazioni con vari gradi di riduzione. Possiamo così distinguere varianti anteriorizzanti (cioè (centro-)anterocentrali): dialetti valenziani ([a̟]), parmigiano 'oltretorrentino' ([ɐ̈]),[2] dialetti bolognesi rustici orientali (Minerbio: [ɐ̈]), o chiaramente anteriori:[3] friulano centrale (cfr. (1)), dialetti catalani nord-occidentali e valenziani ([ɛ]); varianti posteriorizzanti (cioè (centro-)posterocentrali): catalano centrale, comacchiese ([ä̟]), grigionese, ticinese, mantovano ([ä]), o decisamente posteriori (e labiali): varietà carniche (cfr. (1)), catalane nord-occidentali e valenziane ([ɔ]), occitaniche (linguadociano: [ọ], gua-

[1] A questo settore della Romània (cfr. per es. Gsell, 1996; Loporcaro, 2005-06), in cui facciamo rientrare il dominio catalano, nel suo ruolo di 'ponte' tra gallo-romanzo e iberoromanzo (Tagliavini, 1972[6]), appartengono le varietà oggetto d'interesse del presente lavoro. La distinzione tiene conto, tra gli altri fenomeni, proprio del trattamento del vocalismo latino in sillaba finale: cancellazione (nel dominio considerato) *vs.* mantenimento delle vocali finali diverse da -A. Il diverso trattamento riservato al vocalismo non-accentato andrà a sua volta riportato a una diversa struttura della parola nei due blocchi della Romània (cfr. Zamboni, 1990, 1995). Nella sezione nordoccidentale, l'applicazione del processo non è tuttavia priva d'eccezioni: non si allineano a questo sviluppo, per esempio, varietà come il ligure e il veneto centrale e lagunare, in seno all'italo-romanzo settentrionale (cfr. Loporcaro, 2005-06).

[2] La variante anterocentrale, in questa varietà dialettale, può presentare anche una realizzazione più arretrata (fino a [a̠]).

[3] Nella lista delle varietà italo-romanze settentrionali che presentano esiti anteriori(zzanti) potrebbero rientrare anche diversi dialetti di tipo lombardo, che non ci è stato tuttavia possibile esplorare. Ascoli (1873), attingendo a fonti scritte, riporta il caso della Valle d'Intragna (*ibid.*: 255-256), mentre più avanti annota, sulla base delle proprie valutazioni impressionistiche, le località di Tremenico e di Margno in Valsassina (*ibid.*: 502).

scone: [o]); varianti centripete, fino a *schwa*, attraverso vari gradi intermedi: dialetti romagnoli, provenzale, catalano barcellonese e nord-occidentale ([ɐ]), catalano balearico ([ə̟, ɐ]), pavese ([ɐ, ə̠]), antico francese ([ə]). L'ampia gamma di realizzazioni – variamente dislocate nelle fasce medie e semi-bassa del quadrilatero vocalico –, cui ha dato luogo una comune tendenza alla riduzione di -A, comprende dunque varianti periferiche (anteriori e posteriori), varianti anterocentrali, varianti posterocentrali e varianti propriamente centrali. Il grado estremo del processo è rappresentato dal dileguo (francese moderno).[4]

Soggiace dunque a tutte queste realizzazioni un comune processo d'innalzamento, che dovette essere particolarmente attivo (anche se non automatico) in quelle varietà in cui il dileguo di tutte (o quasi) le vocali finali ≠ -A (per l'italo-romanzo settentrionale, cfr. Loporcaro, 2005-06) permise alla vocale bassa d'espandere verso il centro del quadrilatero vocalico, in modo pressoché incontrastato, la propria area di dispersione. Un simile processo dovette avvenire, per esempio, nel gallo-romanzo, in catalano centrale e in varietà friulane; meno intenso dovette presentarsi, invece, in seno al gallo-italico, dove appare piuttosto incipiente.[5]

Tale scenario è anche il più favorevole all'esplicarsi di processi di coarticolazione a distanza tra vocali, se ammettiamo che maggiori sono le dimensioni dell'area d'esistenza delle stesse (come avviene nelle varietà con sistemi ridotti), maggiore è l'intensità con cui i processi in questione possono esplicarsi (Manuel & Krakow, 1984; Bertinetto, 1988; Sánchez Miret, 1999). Proprio una spiegazione di quest'indole riteniamo possa esser invocata per spiegare gli esiti periferici dell'innalzamento, riscontrabili in varietà friulane e catalane (valenziane in particolare), come si mostrerà nei prossimi paragrafi.

---

[4] La panoramica presentata nel § 2 non pretende d'esser esaustiva. Molto utili, per cogliere visivamente la collocazione, nel quadrilatero, delle varianti centralizzanti (e di quelle periferiche) degli esiti di -A, i vocogrammi delle 'fonosintesi' in Canepari (2006^{3}) – relative a svariate lingue, varianti linguistiche e dialetti –, da cui riprendo, per i vantaggi descrittivi che comporta, la distinzione tra vocali 'anteriori', 'anterocentrali', 'centrali', 'posterocentrali' e 'posteriori'. Sulle varietà citate (oltre al friulano e al valenziano, che costituiscono l'oggetto del presente lavoro, e per cui si specificano i riferimenti bibliografici nel corso dell'articolo), si vedano, oltre al citato Canepari: per l'area catalana, Recasens (1991), Veny (1998^{12}), Jiménez (2002), per il pavese, Heilmann (1961), per il bolognese (intramurario e extramurario/rustico/montano), Canepari & Vitali (1995), per il gallo-romanzo in epoca antica, Avalle (2002); una panoramica molto generale degli esiti di -A nelle principali aree romanze si può trovare, tra gli altri, in Lausberg (1976^{2}).

[5] Occorre precisare, infatti, che realizzazioni innalzate ('semi-basse', quali [ɐ, ɐ̈]) non sono ignote all'esterno dei confini del settore della Romània qui considerato: realizzazioni di questo tipo si ritrovano, per esempio, in varietà toscane (e in altre varietà dialettali centro-meridionali), che conservano inalterato, in sillaba finale non-accentata, un sistema ad almeno quattro timbri (/i, e, a, o/). Innalzamenti di quest'entità non sono necesariamente da collegare ai radicali processi di semplificazione del vocalismo d'uscita (: neutralizzazione e cancellazione delle vocali d'uscita diverse da /a/), che hanno interessato gallo-romanzo e gallo-italico, bensì a un naturale 'sconfino' verso una zona 'neutrale' dello spazio vocalico, in cui non sussiste nessun rischio di collisione con gli altri elementi del sottosistema.

3. L'ARMONIA VOCALICA

L'armonia vocalica è un fenomeno assimilatorio che consiste nell'estensione di tutti o d'alcuni tratti d'una vocale ad altri segmenti vocalici, normalmente adiacenti (ma si veda il § 7). L'armonia può manifestarsi in una duplice direzione: da una posizione forte, cioè prominente dal punto di vista percettivo, verso posizioni più deboli; viceversa, da posizioni poco prominenti verso gli altri segmenti vocalici. Nel presente lavoro, si prende in esame il primo dei due modelli d'assimilazione (progressiva, da posizioni forti a posizioni deboli: V → v; cfr. Sánchez Miret, 1999).[6]

4. L'UNIVERSO FRIULANO

Il dominio friulano rappresenta oggi, per dirla col Nievo delle *Confessioni d'un italiano*, un 'piccolo compendio dell'universo' (aggiungiamo: linguistico romanzo nordoccidentale; cfr. § 2), dal momento che vi si può riscontrare, leggibile sul terreno della variazione dialettale, quasi tutta la gamma degli esiti romanzi di -A latina (protoromanza/protofriulana), che permette di seguire le principali tappe della/e trafila/e diacronica/che seguita/e dalla vocale bassa (senza giungere tuttavia al grado estremo del dileguo). Converrà, a tal proposito, distinguere tra: a) posizione prepausale e non-prepausale; b) V *vs.* V + /s/, opposizione che intercetta un importante limite morfologico (nominale: opposizione singolare *vs.* plurale e verbale: 3a pers sing. *vs.* 2a pers. sing.). Entrambi i punti saranno affrontati al § 8.

Lo schema in (1) sintetizza le situazioni prototipiche per V e per V + /s/ in posizione prepausale:[7]

(1)	a. Friul. occidentale comune / goriziano / Ampezzo / cividalese mod. / Beano, Pantianicco, Aquileia, Cervignano	/a/ [a]	/is/
	b. Carnico conservativo / Vito d'Asio / Forgaria, Trasaghis, Osoppo	/a/ [a]	/as/ [as, ɐs][8]

[6] In seno alla Romània nord-occidentale, processi di questo tipo sono documentati, oltre che per le varietà di cui ci occupiamo in questa sede, anche in area ticinese (Salvioni, 1892-94; Sganzini, 1924-26; Loporcaro, 2002; Delucchi, 2008). Tuttavia, nella Svizzera meridionale, -/a/ (già divenuta -/e/; cfr. Loporcaro, *ibid.*) s'assimila (completamente) a qualsiasi vocale presente nella sillaba accentata precedente: [ˈluˑnu] "luna", [ˈteˑre] "tela", [ˈtɛˑrɛ] "terra", [ˈraˑna] "rana", [ˈrɔˑbɔ] "roba", [ˈgoˑro] "gola", [koˈziˑni] "cucina". In questi esempi, relativi alla varietà di Claro (Loporcaro, 2002: 76), abbiamo riportato inalterate le trascrizioni dell'autore; ma cfr. Delucchi (2008: 268-291) sulla neutralizzazione della distinzione tra realizzazioni medio-alte e medio-basse negli stili di elocuzione meno controllati. Invece, in valenziano meridionale – dove -A si conserva generalmente come [a] – e in friulano centro-orientale (escluso il goriziano e altre varietà marginali, dunque solo (1j) e (1k)) – dove -A > [ɛ] – il processo si applica unicamente se la vocale accentata è media e di un grado più alta di quella non-accentata finale: come illustreremo nel § 5, rispettivamente /ˈɛ, ˈɔ/ e /ˈe, ˈo/.

[7] Per una classificazione dei dialetti friulani e una panoramica delle loro principali caratteristiche, si rimanda a Francescato (1966) e Frau (1984).

c. Tramontino (Tr. di Sotto *vs.* di Sopra)	/a/	/as, es/
d. Varietà carniche gortane / Dignano / Sequals, Travesio, S. Quirino (fr. occ.)	/a/	/es/
e. Varietà carniche innovative	/e/	/es/
f. Clauzetto	/e/ [ɛ]	/es/ [ĕs]
g. Treppo Carnico / Paularo	/ə/	/əs/
h. Friul. centrale (udinese) predocument.	/a/?	/es/?
i. Udinese fino XIV sec. ca.	*a* ~ *e* ([ĕ]?)	/is/
j. Varietà centrali conservative / varietà fascia sudor. basso Tagliamento	/e/ [ɛ, ẹ]	/es/ [ĕs, ẹ̆s]
k. Friul. centrale moderno	/e/ [ɛ, ẹ]	/is/
l. Rigolato, Forni Avoltri	/o/ [ɔ]	/os/ [ɔs]
m. Dintorni di Montereale Valcellina / cividalese antico	/o/ [ɔ]	/is/

Per quanto riguarda il valore fonetico della vocale finale nel friulano centrale (udinese) antico (1i), da cui dipende l'attuale -/e/ (innovazione successivamente irradiata a quasi tutta la pianura, la zona collinare centrale, il Medio Tagliamento e le Prealpi orientali – tranne punti conservativi che mantengono -/a/ – e accolta, più recentemente, da diverse varietà carniche innovative), è difficile ricavare sicuri indizi dalle oscillazioni ortografiche presenti nella documentazione scritta di fine Trecento (Vicario, 1999),[9] dove alternano, in modo apparentemente irregolare, le grafie *a* ~ *e*.[10] Un'indicazione ci può venire, da una parte,

[8] [ɐs] sembrerebbe caratterizzare un po' tutto il basso Gorto (Agrons, Muina, Luincis, Ovaro, Entrampo, Cludinico; c.p. di Paolo Roseano). L'eventuale presenza di un'articolazione innalzata in altre aree conservative andrebbe accertata mediante indagini sul campo.

[9] Le fonti documentarie che si rivelano particolarmente utili per seguire l'evoluzione storica del friulano sono soprattutto carte tardomedievali d'uso pratico, che corrispondono alle più diverse tipologie: quaderni di contabilità dei camerari e dei canipari dei comuni e delle confraternite dei maggiori centri delle regione, minute di notai e cancellieri, inventari di beni delle chiese e dei conventi, elenchi di contribuenti, testimonianze e processi, atti di compravendita, ecc. (Vicario, 2001: 13).

[10] Come afferma Vicario (1999: 137), si tratta d'una "continua e imprevedibile alternanza di *-e* e di *-a*, anche per la medesima voce e nei medesimi contesti" che "ci autorizza a ipotizzare che lo scrivente non sentisse la preoccupazione di segnare chiaramente e sistematicamente una *-e* o una *-a* in fine di parola o che non fosse addirittura in grado di farlo". L'interpretazione che dà Vicario di tali oscillazioni grafiche è congruente con la nostra ipotesi: "Il punto centrale della questione è proprio stabilire se la continua confusione di *-e* e di *-a* nella grafia potesse essere dovuto, in qualche modo, alla problematica distinzione di tale fono dal punto di vista percettivo, alla difficoltà di definire con sicurezza il colore della vocale stessa. Se così fosse, si può ipotizzare, per l'udinese del tempo, la presenza di una vocale indistinta di timbro intermedio tra la *a* e la *e*, una vocale che ha sicuramente già risentito del generale indebolimento delle vocali finali latine, una vocale che tende qui ad innal-

dall'esito moderno /e/ e, dall'altra, dalla testimonianza di varietà dialettali (carniche, collinari, prealpine) ancor oggi 'bloccate' in una fase di non completa periferizzazione, ma evidentemente orientate verso la serie anteriore. Possiamo verosimilmente ipotizzare che all'origine del processo d'anteriorizzazione vi sia stato un elevato grado d'instabilità timbrica di -/a/ entro un'ampia area di dispersione, che possiamo immaginare configurata come un'ellisse orientata verso la periferia anteriore del quadrilatero vocalico. L'innalzamento, pertanto, sarebbe stato accompagnato da una progressiva tendenza all'anteriorizzazione, fino alla periferia anteriore del quadrilatero, passando attraverso realizzazioni (centro-)anterocentrali: [e̽] ~ [æ̈] ~ [ɛ̈] ~ [ë]. Fuori dall'area friulana, varianti intermedie non-basse, non più centrali ma neppure ancora completamente anteriori, sono rintracciabili un po' dovunque, come evidenziato nel sintetico panorama dato al § 2 (in area italo-romanza settentrionale, per esempio, le ritroviamo nel parmigiano oltretorrentino e in dialetti bolognesi rustici orientali).

L'ipotesi è che la periferizzazione, ossia l'indentificazione (percettiva e quindi articolatoria) di queste varianti (centro-)anterocentrali di -/a/ con /ˈɛ/ e di quelle simmetriche, posterocentrali, con /ˈɔ/ (se ipotizziamo che il processo d'innalzamento di -A finale latina sia andato di pari passo con una tendenza all'anteriorizzazione o alla posteriorizzazione, a seconda delle zone), possa esser stata favorita dall'attivazione di processi armonici simili a quelli descritti per varietà valenziane meridionali (cfr. § 5.1):[11] come si vedrà, a promuovere l'innalzamento, in una direzione o nell'altra, sono determinate proprietà articolatorie condivise dalla vocale media aperta accentata e dalla vocale finale non-accentata. Inoltre, ciò che prospettiamo per il friulano centro-orientale è che lo stesso processo che ipotizziamo abbia operato in diacronia, portando all'attuale esito -/e/ ([ɛ]), continui a esser attivo in sincronia in alcune varietà (1j,k), dove è responsabile dell'innalzamento di [ɛ] a [e̽] in determinate condizioni contestuali (§ 5.2), con analoghe motivazioni articolatorie e percettive alla base.

L'universo variegato delle soluzioni offerte dai dialetti a sud di Valenza rende manifeste, proiettate in sincronia sullo spazio diatopico, le diverse tappe ipotizzabili del percorso evolutivo che dovette condurre all'attuale assetto friulano. A render ancor più evidente l'affinità tra valenziano e friulano, per quanto riguarda i processi descritti, è il fatto che anche in area valenziana si riscontrano condizioni di variabilità geofonica molto simili a quelle friulane: dialetti con completa periferizzazione convivono accanto ad altri che sembrano essersi arrestati a tappe intermedie (è il caso, per esempio, di Sueca (§ 5.1), che presenta [e̽] invece di [ɛ]).

---

zarsi verso la serie anteriore [...] ma che non ha ancora raggiunto un grado di altezza tale da renderla chiaramente una *-e*, immediatamente distinguibile da una *-a*" (*ibid.*).

[11] Quanto prospettiamo per il friulano non deve necessariamente valere anche per altre varietà romanze che presentano varianti periferiche – anteriori o posteriori – per -A, le quali possono esser dovute, piuttosto, a dinamiche strutturali, interne al sistema (propulsione, trazione, ecc.). La variante postero-labiata media che caratterizza dialetti occitanici (§ 2), per esempio, sembra esser il risultato del riempimento, da parte della vocale bassa, d'una casella lasciata vuota in seguito a una serie di mutamenti a catena innescati dall'anteriorizzazione di /u/ e dal conseguente innalzamento di /o/ (sia accentata che non-accentata), la cui casella poté esser occupata, a sua volta (a partire dal XV-XVI sec.), in una vasta area, da /a/ (Lafont, 1991: 4).

5. L'ARMONIA VOCALICA IN VALENZIANO E FRIULANO

Valenziano e friulano sono accomunati da un sistema eptavocalico d'elementi brevi (/i, e, ɛ, a, ɔ, o, u/),[12] che oppone, in posizione accentata, due medie chiuse ([+ATR]):[13] /e, o/ a due aperte ([-ATR]): /ɛ, ɔ/. In posizione non-accentata, il valenziano distingue cinque vocali d'uscita (/i, e, a, o, u/; tuttavia /o/ è molto raro), il friulano solo tre (/i, e, o/; anche in questo caso, l'ultimo compare in pochissimi esempi, perlopiù d'origine veneta e italiana).[14] Il presente paragrafo prende in esame gli effetti dell'assimilazione a distanza tra le

[12] La maggior parte delle varietà friulane centro-orientali e carniche utilizza la durata con funzione distintiva (Francescato, 1966; Frau, 1984; Vanelli, 1998; Miotti, 2002). Il sistema più diffuso oppone sette elementi brevi a cinque elementi lunghi (/ɛː, ɔː/ sono presenti solo in varietà marginali).

[13] Il tratto [ATR] (*Advanced Tongue Root*, 'radice della lingua avanzata', in riferimento allo spostamento in avanti della radice della lingua, con conseguente allargamento della cavità faringea, spesso accompagnato da innalzamento della lingua; cfr. Nespor, 1993: 57) è stato sempre più usato, soprattutto negli ultimi anni, in sostituzione della tradizionale opposizione 'teso' – 'rilassato' (per una verifica sull'equivalenza fra i due tipi di tratti, cfr. Ladefoged & Maddieson, 1996: 300-306). Per l'ambito romanzo, il tratto [±ATR] ha assunto un ruolo sempre più centrale, per quanto riguarda l'analisi delle armonie vocaliche e dei processi metafonetici, grazie a Calabrese, che lo utilizza definitivamente, in sostituzione di [±teso], a partire dalla fine degli anni Ottanta (cfr. Calabrese, 1989; Grimaldi, 2003).

[14] Normalmente, le vocali finali diverse da -/a/ dileguano in entrambe le varietà. Se escludiamo -/o/, rarissimo in valenziano, e in friulano limitato a pochissimi prestiti, per quanto perfettamente acclimatati nel vocabolario comune, le altre vocali d'uscita non sono tutte etimologiche; si tratta infatti o del risultato della risoluzione di sequenze etimologiche particolari (CjV, CwV; cfr. val. -/i/, -/u/: *idoni* 'idoneo', *assidu* 'assiduo'; friul. *lunari* 'lunario' e altri prestiti posmedievali, di contro all'esito regolare -/aːr/ di lat. -ARIU) oppure di vocali d'appoggio, apparse in fase romanza. Sono di quest'ultimo tipo -/e/, in valenziano, e -/i/, in friulano, con analoghe caratteristiche (elementi morfologicamente non marcati) e analoga funzione e distribuzione. In entrambe le varietà sono richieste, infatti, dopo nessi consonantici, risolti o no: val. /ˈpaɾe/, friul. /ˈpa(ː)ri/ 'padre', val. /ˈnɔstɾe/, friul. /ˈnestri/ 'nostro', oppure come vocale d'appoggio nella prima persona sing. del presente indicativo (e nel congiuntivo): val. /ˈkante/, friul. /ˈcanti/ 'canto', laddove l'esito regolare (attestato sia in valenziano che in friulano antico) avrebbe dovuto avere desinenza Ø < lat. CANTO, per effetto, come già detto, del dileguo delle vocali finali dieverse da -/a/. L'aggiunta, in questo caso, può esser spiegata come strategia per sanare un'asimmetria sillabica tra le prima e la seconda persona, che aveva una sillaba in più (cfr., per il friulano, Benincà & Vanelli, 1975). In friulano si ha -/i/, inizialmente per anaptissi, anche nella sillaba postaccentata di proparossitoni latini, poi rimasta scoperta in seguito al dileguo della C finale: /ˈmjedi/ 'medico', /sˈtɔmi/ 'stomaco', /ˈbɛvi/ 'bere' (rispettivam. < lat. MEDICU, STOMACU, BIBERE). Si ha -/i/- come vocale d'appoggio, più che come risultato dell'innalzamento della vocale postaccentata (cfr. nota 31), anche in casi come /ˈd͡ʒɔvin/ 'giovane', /ˈwarfin/ 'orfano' (lat. IUVENE, ORPHANU); la prova sarebbe l'esistenza d'esempi come /ˈmangin/ 'mangano' (lat. MANGANU), in cui l'anaptissi di /i/ non ha provocato l'intacco della consonante precedente (come sarebbe avvenuto se l'elemento vocalico fosse stato etimologico, come in /ˈward͡ʒine/ 'aratro' < lat. ORGANU; cfr. nota 31 e Benincà, 1989: 568). Infine, troviamo -

medie accentate e i riflessi attuali di -A latina nelle rispettive varietà (rispettivamente, -/a/ in valenziano e -/e/ in friulano centrale). In particolare, i § 5.1 e 5.2 mostrano, nel dettaglio, le modalità con cui si esplica, in sincronia, il condizionamento coarticolatorio nei contesti armonici nelle due aree romanze, mentre il § 5.3 propone una ricostruzione in chiave diacronica a partire dai dati disponibili in sincronia.

5.1 L'armonia vocalica nel valenziano meridionale

- Modello prototipico (Canals): /ˈɛ, ˈɔ/ innalzano /a/ finale diffondendo i rispettivi tratti 'punto d'articolazione', rispettivamente 'palatale' e 'labiale': /a/ → [ɛ]; /a/ → [ɔ]. In altre parole, -/a/ si assimila completamente alla vocale semiaperta[15] accentata precedente, come esemplificato in (2) e (3).[16] Nei contesti non-armonici, -/a/ si mantiene invece come [a].

(2)

(3)

a.	/ˈtɛla/	[ˈtɛlɛ]	*tela*	"tela"	b.	/ˈkapa/	[ˈkapa]	*capa*	"cappa"
	/ˈpɔɾta/	[ˈpɔɾtɔ]	*porta*	"porta"		/ˈpeɾa/	[ˈpeɾa]	*pera*	"pera"
						/ˈmiɾa/	[ˈmiɾa]	*mira*	"guarda"
						/ˈtota/	[ˈtota]	*tota*	"tutta"
						/ˈʎuna/	[ˈʎuna]	*lluna*	"luna"

- Varietà restrittive: /a/ finale si assimila solo a /ɛ/ (Cullera) o solo a /ɔ/ (Borriana). I rimanenti contesti, non-armonici, mantengono /a/ come [a].

- Alcune varietà hanno esteso il prodotto dell'armonia anche ai contesti non-armonici, generalizzando la realizzazione anteriore (per es. Sueca, che, come già detto, presenta [e̯] piuttosto che [ɛ]) o quella posteriore (per es. Xaló); nei casi in cui a diffondere il tratto di punto d'articolazione siano sia /ˈɛ/ che /ˈɔ/, la neutralizzazione di /a/ nei contesti non-armonici può avvenire sia con la variante anteriore (Benitatxell) che con quella posteriore (Ontinyent):

(4)

a. Sueca: [ˈtɛle̯] = [ˈpɔɾte̯, ˈkape̯, ˈmiɾe̯, ˈʎune̯]
b. Xaló: [ˈpɔɾtɔ] = [ˈtɛlɔ, ˈkapɔ, ˈmiɾɔ, ˈʎunɔ]
c. Benitatxell: [ˈtɛlɛ] = [ˈkapɛ, ˈmiɾɛ, ˈʎunɛ] *vs.* [ˈpɔɾtɔ]
d. Ontinyent: [ˈtɛlɛ] *vs.* [ˈpɔɾtɔ] = [ˈkapɔ, ˈmiɾɔ, ˈʎunɔ]

/i/, -/u/ come secondo elemento di dittonghi di varia origine: val./friul. /ˈmai/ 'mai', friul. /caˈvai/ 'cavalli', val. /ˈpau/ 'pace'.

[15] In quest'articolo, 'semiaperto' e 'semichiuso' vengono utilizzati come traduzione di *open-mid* e *close-mid*, rispettivamente.

[16] Tutti gli esempi valenziani utilizzati nel presente lavoro sono tratti dai lavori di Jiménez e Jiménez & Lloret citati nel corso dell'articolo e testati con l'aiuto di tre informanti nativi, rappresentativi dei diversi modelli d'armonia qui discussi.

5.2 L'armonia vocalica nel friulano centro-orientale[17]

- /e/ finale (< -A latina) è realizzata come semiaperta [ɛ] nei contesti non-armonici (7b),[18] esito d'un processo d'innalzamento/avanzamento di -/a/, stabilizzatosi, nel friulano centrale (udinese), nel corso del XV sec.

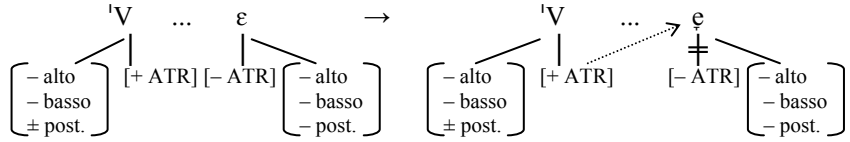
- /ˈe, ˈo/ [ẹ, ọ] innalzano [ɛ] finale fino a [ẹ], che conserva il punto d'articolazione palatale, che non viene percettivamente compromesso dalla pur presente tendenza – fisiologica (e statisticamente significativa nella maggior parte dei casi, sopprattuto per il soggetto maschile) – alla posteriorizzazione dopo vocale posteriore (cfr. Fig. 1 e Fig. 2). A parità di tutti gli altri tratti, a esser diffuso sulla vocale finale è dunque il tratto [+ATR] della vocale media semichiusa accentata, come esemplificato in (5) e (6):[19]

[17] Preferiamo mantenere, per comodità, la denominazione generica di 'friulano centro-orientale', o semplicemente 'centrale', sebbene i risultati dell'indagine sperimentale che presentiamo in questa sede siano in realtà relativi alla sua sottovarietà più occidentale, quella della cosiddetta 'fascia sudorientale del basso Tagliamento' (cfr. (1j); Frau, 1984; Miotti, 2007). Future indagini più approfondite saranno condotte al fine d'accertare le modalità con cui si manifestano i processi qui discussi nelle aree propriamente centrali e orientali.

[18] In realtà, anche nella categoria '[ɛ]', rappresentata dai contesti non-armonici, potrebbero venir riconosciute ulteriori sottocategorie, in base al condizionamento esercitato dalla vocale accentata. Ciò è particolarmente evidente nel soggetto femminile di Camino al Tagliamento (cfr. Tab. 2), in cui sembra esser attivo, stavolta, il processo opposto a quello dell'assimilazione a distanza, cioè un fenomeno di dissimilazione timbrica (a distanza), innescato dalla presenza d'una vocale alta in sillaba accentata, per cui -/e/ risulta abbassarsi ulteriormente dopo /ˈi, ˈu/ (il relativo, più o meno sensibile, innalzamento di -[ɛ] dopo la vocale con grado d'apertura massimale (/ˈa/) è invece decisamente irregolare in tutti soggetti: nel soggetto maschile, per esempio, è significativo solo dopo dentale). Fenomeni di dissimilazione timbrica a distanza, innescati dalla presenza d'una vocale alta all'interno della sillaba prominente, non sono sconosciuti in area italo-romanza. Anche in italiano neutro si ha /ˈtɔpo/ [ˈtɔːpo] ma /ˈtipo/ [ˈtiːpọ] (Canepari, 1999[2]: 59), ma in veneziano e in trevigiano (cittadini, sia in dialetto che in italiano regionale), e in veneto-giuliano, si può arrivare addirittura a [ˈtiːpɔ] (*ibid.*: 398, 406). La dissimilazione riscontrata per il friulano potrà dunque esser messa in relazione con tendenze italo-romanze (o almeno con quelle documentate per i contigui dialetti veneti). È prematuro concludere se tale tendenza, peraltro regolare nel soggetto femminile (qualunque sia la consonante coronale che precede la vocale finale), sia direttamente correlabile a fattori quali il sesso del parlante o il dialetto (i due informanti, maschile e femminile, di cui si comparano i dati, sono parlanti di varietà diverse, pur all'interno della stessa area dialettale). Si tratta evidentemente di conclusioni provvisorie, da verificare con un campione più ampio di soggetti.

[19] L'indagine è stata circoscritta al contesto 'vocale finale preceduta da consonante coronale'. L'effetto del tipo di consonante ([+cor] *vs* [-cor]) sull'esplicazione dei processi armonici sarà oggetto d'un'indagine specifica. Le varie figure e tabelle inserite nel presente lavoro sono relative a due dei quattro soggetti friulani utilizzati per l'indagine acustica e han-

(5)



(6)

a.	/ˈkwete/	[ˈkwe̯te̯]	*cuete*	“cotta’	b. /ˈmate/	[ˈmatɛ]	*mate*	“matta”
	/ˈmote/	[ˈmo̯te̯]	*mote*	“mossa”	/ˈsɛde/	[ˈsɛe̯dɛ]	*sede*	“seta”
					/siˈzile/	[siˈziilɛ]	*sisile*	“rondine”
					/ˈkɔde/	[ˈkɔo̯dɛ]	*code*	“coda”
					/ˈvude/	[ˈvuudɛ]	*vude*	“avuta”

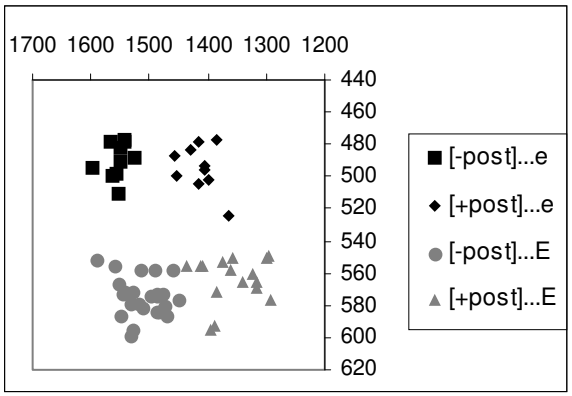
La figura 1 mostra il condizionamento coarticolatorio -/e/ preceduta da C vibrante in contesti armonici (‘...e’) *vs.* non-armonici (‘...E’), dopo V anteriori *vs.* posteriori (soggetto maschile).[20] I valori medi e relative deviazioni standard sono riportati a destra.[21]

---

no il semplice scopo d’esemplificare e supportare, mediante evidenze sperimentali, i fenomeni discussi per quanto riguarda il friulano.

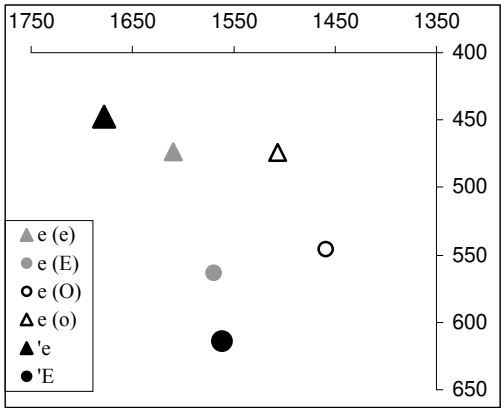
[20] Abbiamo incluso tra le [-post] anche /a/, dal momento che i valori rilevati per F2 non sono significativamente diversi da quelli calcolati per le vocali [-post] propriamente dette, mentre lo sono se confrontati con quelli delle [+post].

[21] Per quanto concerne l’indagine acustica, il protocollo sperimentale è stato il seguente: le parole bersaglio sono state inserite in posizione finale (tonìa) e in posizione interna di frase cornice; sono state richieste almeno cinque ripetizioni di ciascuna struttura. I valori formantici (F0, F1, F2 e F3, in Hz) delle vocali in sillaba accentata e non-accentate finali sono stati misurati, mediante *Praat*, in corrispondenza della porzione centrale del segmento vocalico, in una varietà di contesti consonantici e proiettati poi su piani cartesiani di tipo tradizionale. Nel corpus sono presenti le seguenti consonanti (coronali) prevocaliche: /t, d, s, z, r, l/. Gli informanti sottoposti all’indagine acustica sono quattro soggetti friulani – tre di sesso femminile e uno di sesso maschile –, tutti parlanti nativi della varietà friulana parlata nelle rispettive località (appartenenti alla sottovarietà di friulano centro-orientale nota come ‘fascia sudorientale del basso Tagliamento’), e tre soggetti valenziani, uno si sesso maschile e due di sesso femminile (di varie località a sud di Valenza: rispettivamente, Olleria, Sueca e Carcaixent), anch’essi parlanti nativi delle rispettive varietà.



	F1	F2
[-post]...e	490 (11)	1554
[+post]...e	495 (14)	(19)
[-post]...E	575 (12)	1413
[+post]...E	564 (14)	(29)
		1511
		(35)
		1357
		(46)

Figura 1: Diagramma di dispersione relativo a -/e/ preceduta da C vibrante in friulano (simboli SAMPA)



	F1	F2
(‘e)...e	474	1609
(‘E)...E	563	1570
(‘o)...e	474	1507
(‘O)...E	545	1459
‘e	447	1678
‘E	614	1562

Figura 2: -/e/ preceduta da C dentale in friulano (valori medi): condizionamento coarticolatorio dopo V medie (tra parentesi).
Simboli SAMPA (v. Tabella 1 per dettagli)

Soggetto maschile, Sedegliano (UD) (-V dopo C dentale)								
	F1 media	F1 min	F1max	F1 ds	F2 media	F2 min	F2 max	F2 ds
ˈe	447	412	490	19	1678	1414	1853	111
ˈɛ	614	548	658	33	1562	1393	1636	58
ˈo	469	442	493	20	901	833	974	46
ˈɔ	575	518	623	28	989	889	1042	42
(ˈe)...ẹ	474	434	518	18	1609	1527	1716	44
(ˈo)...ẹ	474	453	496	14	1507	1457	1544	28
(ˈa)...ɛ	535	518	559	12	1575	1514	1613	30
(ˈɛ)...ɛ	563	536	619	21	1570	1519	1622	25
(ˈi)...ɛ	552	529	583	18	1554	1461	1622	51
(ˈɔ)...ɛ	545	515	579	19	1459	1401	1527	38
(ˈu)...ɛ	548	523	585	17	1410	1302	1482	48
media -ẹ	474	434	518	17	1588	1457	1716	59
media -ɛ	549	515	619	20	1511	1302	1622	78

Tabella 1: Valori medi e indici statistici (valore minimo, massimo e deviazione standard) di F1 e F2, relativi alle V medie accentate e a -/e/ non-accentata finale (soggetto maschile)

Soggetto femminile, Camino al Tagliamento (UD) (-V dopo C vibrante)								
	F1 media	F1 min	F1max	F1 ds	F2 media	F2 min	F2 max	F2 ds
ˈe	516	460	572	27	2346	2211	2514	124
ˈɛ	744	701	798	25	1930	1855	2019	44
ˈo	501	421	552	34	845	761	933	45
ˈɔ	652	578	781	73	910	850	958	39
(ˈe)...ẹ	598	549	678	33	1915	1816	1976	44
(ˈo)...ẹ	590	530	635	24	1843	1791	1891	31
(ˈa)...ɛ	654	619	696	28	1830	1792	1863	31
(ˈɛ)...ɛ	692	648	736	27	1796	1738	1881	39
(ˈi)...ɛ	746	697	789	25	1793	1706	1849	45
(ˈɔ)...ɛ	695	651	755	36	1779	1698	1837	48
(ˈu)...ɛ	755	720	782	25	1746	1684	1796	41
media -ẹ	594	530	678	29	1879	1791	1976	52
media -ɛ	707	619	789	42	1787	1684	1863	46

Tabella : Valori medi e indici statistici (valore minimo, massimo e deviazione standard) di F1 e F2, relativi alle V medie accentate e a -/e/ non-accentata finale (soggetto femminile)

5.3 Discussione

È ipotizzabile che in friulano, se teniamo conto degli indizi cui s'è accennato nel § 4, il processo d'innalzamento di -A si sia innescato – già accompagnato comunque da una chiara tendenza all'avanzamento o all'arretramento – in una fase precedente la vera e propria periferizzazione verso la serie anteriore o posteriore. Solo in seguito si sarebbe avuta l'assimilazione completa: le vocali medie aperte accentate avrebbero attirato a sé la non-accentata finale (già innalzata e articolatoriamente instabile), diffondendo il tratto 'punto d'articolazione' (palatale o labiale). Il grado di maggior anteriorità o posteriorità della variante centralizzata finale di -A, secondo una chiara distribuzione diatopicamente complementare (tendenza all'anteriorizzazione in pianura *vs.* tendenza alla posteriorizzazione nel cividalese e nell'Alto Gorto),[22] dovette determinare i due principali esiti -/e/ [ɛ] e -/o/ [ɔ], poi generalizzatisi, per analogia, anche ai contesti non-armonici (come avvenuto nelle varietà valenziane in (4a) e (4b)).

Cruciale, nell'applicazione del processo, sia per l'area iberica che per l'area friulana, oltre al già menzionato carattere intrinsecamente instabile della vocale non-accentata finale, la vicinanza articolatoria tra la vocale in posizione forte e la vocale in posizione debole (la differenza è d'un solo grado d'altezza), che possono esser mantenute adeguatamente differenziate solo a costo d'un notevole sforzo (in termini di numero di gesti articolatori) da parte del parlante. La prossimità articolatoria, come fattore che favorisce la confusione e l'assimilazione dei segmenti vocalici coinvolti nei processi descritti, continua a esser decisiva nel friulano centrale moderno (tra /ˈe/ e l'esito [ɛ] di lat. -A) così come dovette esserlo, come ipotizziamo, nell'udinese tardomedievale (tra /ˈɛ/ e le realizzazioni centralizzate e anteriorizzate di -/a/), come esemplificato in (7):

[22] Va detto che -/o/, che non prendiamo in considerazione in questa sede, sta oggi regredendo, per tornare a -/a/, mentre -/e/ prende sempre più piede. È probabile che, in origine, -/a/ fosse velarizzata ([a̠ ~ ɑ ~ ɐ̟ ~ ɐ̈ ~ ɔ]) in tutto o gran parte del Friuli settentrionale (oscillazioni *a* ~ *o* sono attestate anche in testi antichi riferibili a Gemona e ad Artegna; cfr. Pellegrini, 1994), lungo un arco che, grosso modo, doveva unire Cividale (a est di Udine) alla Carnia centro(-orientale), fino a Montereale Valcellina, nel Friuli nord-occidentale (Castellani, 1980: 52-55), mentre nell'udinese prevalessero realizzazioni anteriorizzanti ([a̟ ~ ɐ̟ ~ ë]). La variante udinese, collegata a un modello linguistico più prestigioso, avrebbe finito col render stigmatizzate le realizzazioni posteriorizzanti o già del tutto assimilate alla serie posteriore ([ɔ]), facendole regredire a [a]. Il mutamento sembra esser tuttora in corso, se è vero che, per citare un esempio, Gortani (in Marinelli & Gortani, 1924-25) attestava -/o/ in località, come Ravascletto (Alto Gorto), che oggi (in realtà già negli anni Sessanta, come risulta in Francescato, 1966: 407) non ne serbano più traccia. Il cedimento sembra favorito dalla posizione davanti a /s/, se è vero che persino Rigolato, una delle roccaforti di -/o/, davanti a sibilante preferisce /e/ (esattamente come a Ravascletto, che ora ha -/a/ *vs.* -/es/).

(7)	Lat.	*nit(i)da*	*cocta*
	Protofriul.	[ˈnɛta]	[ˈkwɐ̣ta]
	Udinese fino a XV sec.	[ˈnɛtɐ̯̣ ~ ɐ̯̣̈]	[ˈkwɐ̣tɐ̯̣ ~ ɐ̯̣̈]
		[ˈnɛtɛ]	[ˈkwɐ̣tɛ]
	Friul. centr. moderno	[ˈnɛtɛ]	[ˈkwɐ̣tɐ̣]

L'effetto coarticolatorio avrebbe comportato benefici di natura articolatoria (cfr. § 7), dal momento che lo sforzo di mantenere differenziate due vocali è tanto maggiore quanto più ravvicinate sono nello spazio articolatorio.[23] La prossimità articolatoria tra il timbro della vocale in posizione forte e quello della vocale in posizione debole risulta accentuata, nelle varietà valenziane in cui si produce il fenomeno (rispetto ad altre in cui invece non si applica), dall'estrema apertura delle medie aperte che caratterizza tali varietà (Recasens, 1991: 100; Jiménez, 2001). Per quel che concerne il friulano, diacronicamente, avrebbero favorito l'assimilazione di -A alle medie aperte (-A > [ɛ] o [ɔ], con la distribuzione diatopica vista): a) l'apprezzabile innalzamento, accompagnato da considerevole dispersione verso la zona anterocentrale (in pianura) o verso la zona posterocentrale del quadrilatero (a Cividale e nell'Alto Gorto), della vocale finale (processo innescatosi successivamente al dileguo delle vocali finali ≠ /a/), che ne favorì l'avvicinamento articolatorio; b) probabilmente, come fattore coadiuvante, la realizzazione più bassa di /ˈɛ, ˈɔ/ nelle varietà pianigiane rispetto a quelle carniche più conservative (dove la distinzione sul piano fonetico, almeno per /ˈe/ ~ /ˈɛ/, è in generale meno evidente).[24]

---

[23] D'altra parte, come propone Jiménez (2001: 232-233), "si analitzem la qüestió des del punt de vista de les vocals que desencadenen el procés, es pot observar que /ɛ/ i /ɔ/ són les vocals marcades del sistema: poden aparèixer en la posició més perceptible –la posició tònica–, però no en la posició menys prominent –la posició àtona–, on se suposa que només es mantenen els trets no marcats. Per tant, sembla esperable que hi hagi fenòmens que tractin de preservar les vocals obertes i, fins i tot, reforçar-les en posicions prominents (Cole & Kissenberth 1995, Jiménez 1998). L'extensió cap a les vocals veïnes seria una manera d'assegurar la percepció de les vocals obertes". Nel caso specifico del friulano, si potrebbe anche avanzar l'ipotesi, data la relativa vicinanza articolatoria tra medie semichiuse e medie semiaperte, e in particolare la realizzazione abbassata delle prime, che la diffusione di tratti alla vocale finale abbia la funzione di render percettivamente più stabile l'opposizione /e/ ~ /ɛ/.

[24] È interessante rilevare che, nel caso di /ˈɛ/, come pure per /ˈe/ (cfr. nota seguente), per il friulano (pianigiano) abbiamo riscontrato valori di F1 relativamente superiori rispetto alla media (almeno in ambito italo-romanzo), il che rende ancor più visibile il parallelismo tra varietà friulane e valenziane, per quanto riguarda le cause che hanno attivato i processi armonici. Si confronti (facendo astrazione dalla varietà di condizionamenti, da quelli contestuali a quelli metodologici fino a quelli legati alle condizioni sperimentali, che rendono solo parzialmente confrontabili i valori assoluti) il valore medio di F1, relativo al nostro soggetto maschile (614 Hz; cfr. Tab. 1), con quelli calcolati per l'italiano settentrionale (557 Hz, in Ferrero, 1979), da Albano Leoni & Maturi (1998[2]: 102), per l'italiano televisivo (500 Hz) e con quelli tabulati in Sorianello (2002) per l'italiano senese, fiorentino e livornese (rispettivamente 479, 490 e 562 Hz). Solo l'italiano pisano presenta valori comparabili

Sincronicamente, l'ulteriore innalzamento di -/e/ dopo /ˈe, ˈo/, peraltro non obbligatorio (per es., la varietà di Clauzetto (1f), nel Friuli nord-occidentale, che ha -A > -/e/, è ferma allo stadio -[ɛ], per tutti i contesti), può esser spiegato in modo analogo: la relativa vicinanza articolatoria tra la media semichiusa accentata (di fatto [ę]) e [ɛ] finale non-accentata (per quanto i rispettivi timbri siano uditivamente e acusticamente ben differenziati; cfr. Tab. 1 e 2)[25] può aver favorito il processo coarticolatorio.

6. IL DOMINIO DELL'ARMONIA VOCALICA

Oltre ai casi in (6a), in cui il tratto [+ATR] della vocale accentata viene diffuso sulla vocale finale contigua, il friulano centrale presenta armonia anche nelle situazioni esemplificate in (9a), da contrastare con (9b):[26]

---

con quelli friulani (628 Hz; cfr. Calamai, 2002). Per il friulano, nostre indagini precedenti avevano confermato valori simili a quelli del soggetto qui esaminato (di poco superiori o inferiori ai 600 Hz) anche per vari parlanti d'area centrale.

[25] Nel friulano centro-orientale, la vicinanza articolatoria tra /ˈe/ e [ɛ] finale non-accentata è accentuata da una realizzazione di /ˈe/ relativamente più bassa rispetto ai valori medi tipici riportati in letteratura per il vocoide semichiuso. Per quanto concerne F1 (correlato acustico dell'altezza), tali valori, normalmente, rimangono compresi, in soggetti maschili, tra i 350 e i 400 Hz. Si confrontino, per esempio, i valori friulani dati nella Tab. 1 con i 400 Hz dell'italiano settentrionale (in Ferrero, 1979), i 375 Hz relativi all'italiano parlato naturale d'un campione di giornalisti Rai provenienti da varie regioni (Albano Leoni & Maturi, 1998[2]: 102), i 397 Hz dell'italiano pisano (liste di parole) misurati da Calamai (2002) e i 384, 350 e 391 Hz (dialoghi strutturati) dell'italiano rispettivamente senese, fiorentino e livornese, riportati in Sorianello (2002).

[26] Diffusione di tratti, in friulano, si ha anche verso sinistra, almeno fino alla vocale contigua a quella accentata, sebbene, acusticamente e percettivamente, le proporzioni del fenomeno non siano così vistose come quelle viste per la direzione 'canonica' (verso destra). Casi come /maleˈdet/ [ˌmaleˈdẹt ~ -lẹ-] 'maledetto' *vs* /maleˈdɛte/ [ˌmalɛ̨ˈdɛtɛ ~ -lẹ-] 'maledetta' mostrano che, stavolta, in posizione non-accentata non-finale, dove normalmente i quattro gradi d'apertura del vocalismo accentato si riducono a tre (con neutralizzazione in /e, o/ del grado d'apertura delle medie), sono /ɛ, ɔ/ accentate a diffondere il tratto [-ATR] sulle medie precedenti. Si tratta in realtà, alla luce dei risultati provvisori della nostra indagine, d'una tendenza, più che d'un fatto sistematico: i valori minimi di F1 (correlato acustico del grado di sollevamento/abbassamento del dorso della lingua) sembrano infatti coincidere per entrambe le realizzazioni ('normale' *vs* abbassata), mentre quelli massimi sono decisamente più alti per la variante più aperta che, pertanto, è caratterizzata da una maggior dispersione. Riportiamo, a titolo esemplificativo, i dati relativi a F1 (valori in Hz) per due soggetti, uno di sesso maschile: variante 'normale' (media: 475, min: 464, max: 501, ds: 12), variante abbassata (media: 514, min: 472, max: 556, ds: 30), e uno di sesso femminile: variante 'normale' (media: 478, min: 465, max: 510, ds: 14), variante abbassata (media: 521, min: 463, max: 581, ds: 44). Sebbene però l'area di dispersione della realizzazione bassa inglobi quella della variante 'normale' in tutti gl'informanti esaminati, la differenza tra le due categorie è sempre risultata significativa al *t* test ($p < 0,05$). La questione è di grande importanza per la comprensione dei processi che regolano l'armonia vocalica in friulano, e merita certamente un'indagine più approfondita.

(8) a. /ˈjevile/ [ˈjẹvilẹ] *jevile* “alzala”
/ˈmovile/ [ˈmọvilẹ] *movile* “muovila”
b. /ˈlavile/ [ˈlavilɛ] *lavile* “‘avala”
/ˈpɛlile/ [ˈpɛlilɛ] *pelile* “pelala”
/ˈsintile/ [ˈsintilɛ] *sintile* “sentila”
/ˈkɔpile/ [ˈkɔpilɛ] *copile* “uccidila”
/ˈbutile/ [ˈbutilɛ] *butile* “buttala”

Non sembrano influire nell’estensione del dominio d’applicazione dell’armonia vocalica le restrizioni morfologiche che, al contrario, ritroviamo operanti nel valenziano prototipico:

(9) a. /ˈpɛlala/ [ˈpɛlɛla] *-[lɛ] *pelala* “pelala”
/ˈpɔɾtala/ [ˈpɔɾtɔla] *-[lɔ] *portala* “portala”
b. /ˈpɛɾla/ [ˈpɛɾla] *-[lɛ] *perdla* “perdila” (ma /ˈpɛɾla/ [ˈpɛɾlɛ] *perla* “perla”)
/ˈkɔula/ [ˈkɔula] *-[lɔ] *coula* “cuocila”

Gli esempi elencati in (8) e (9) sollevano il problema della determinazione dei fattori – fonologici o morfolessicali – che intervengono nel definire l’armonia vocalica: si tratta di determinare se i clitici vengano aggiunti al loro verbo ospite a un livello più alto della gerarchia prosodica, costituendo così una parola fonologica separata (cfr. per es. Nespor & Vogel, 1986), oppure se formino col loro verbo ospite un’unica parola fonologica postlessicale (Dressler, 1985; Booij, 1996; Loporcaro, 1999, 2000, 2002). In valenziano, sembra fuori discussione che il dominio massimo dell’armonia vocalica sia la parola fonologica: le vocali dei pronomi clitici non vengono coinvolte nel processo (Jiménez, 2001: 229-230), neppure se la vocale del clitico è adiacente a quella accentata (9b). È evidente, dunque, la presenza d’un’importante restrizione di carattere morfologico (morfolessicale). La contiguità del contesto armonico è importante, in quanto delimita l’ambito prosodico (metrico) dell’armonia vocalica, che in valenziano sembra essere il piede metrico (binario), in cui la posizione debole si trova a destra di quella forte, da cui vengono diffusi i tratti armonici (Jiménez, 2001, 2002). Coerentemente, vengono esclusi dall’applicazione del processo armonico casi come quelli riportati in (10):

(10) /ˈtɛtɾika/ [ˈtɛtɾika] (*-[kɛ]) *tètrica* “tetra”
/ˈrːɔtula/ [ˈrːɔtula] (*-[lɔ]) *ròtula* “rotula”
/ˈpɛkoɾa/ [ˈpɛkoɾa] (*-[ɾɛ]) *pècora* “pecora (insulto)”

in cui la diffusione del tratto ‘punto d’articolazione’ è bloccata dalla presenza d’una vocale intermedia tra la media accentata e la vocale bersaglio (/a/) finale. Per il friulano, invece, è evidente che il dominio dell’armonia è la parola fonologica postlessicale: i clitici ricadono all’interno della parola fonologica a livello postlessicale; il dominio dell’armonia vocalica è pertanto fonologicamente definito. Metricamente, i clitici ricadono all’interno dello stesso

piede, che questa volta questa volta dobbiamo postulare ternario[27], in cui si trova la vocale che scatena il fenomeno armonico (esattamente come nel dialetto ticinese di Claro, studiato da Loporcaro, 2002).

7. LE CAUSE DELL'ARMONIA VOCALICA: ESIGENZE ARTICOLATORIE O PERCETTIVE?

I processi d'assimilazione a distanza sono riportati, in letteratura, al conseguimento di benefici di varia natura, che possiamo così riassumere: a) semplificazione articolatoria (il gesto coarticolatorio risulta semplificato ed economico; cfr., per es., Pulleyblank, 2002), b) benefici percettivi (cfr. Walker, 2005, 2006), c) entrambi (cfr. Cole & Kisseberth, 1994).[28]

Circa le esigenze, percettive oppure articolatorie, a cui rispondono i modelli d'armonia vocalica visti per valenziano e friulano (§ 5), è fuor di dubbio che la semplificazione articolatoria, in termini di riduzione dei gesti coarticolatori (o d'inerzia degli articolatori: Pulleyblank, 2002) che presuppongono, all'interno dei rispettivi domini, i processi di armonia vocalica,[29] parla a favore d'un'interpretazione articolatoria, piuttosto che percettiva dei processi discussi. Mentre un'interpretazione percettiva è generalmente adeguata in quei casi in cui i tratti armonici si diffondono da posizioni deboli verso elementi più prominenti, allo scopo d'aumentarne la percettibilità – ma senza che ciò comporti, necessariamente, un beneficio articolatorio (Walker, 2005, 2006; Jiménez & Lloret, c.d.s.) –, nei casi considerati la diffusione di tratti avviene a partire da elementi *già* prominenti. Poiché, verosimilmente, la semplificazione articolatoria coinvolge elementi adiacenti, ci si aspetta che il dominio dell'armonia sia omogeneo. Mentre le varietà valenziane non pongono nessun tipo di problema, da questo punto di vista, i casi friulani in (8) sembrerebbero invece smentire l'ipotesi articolatoria. Tuttavia, se si tiene presente la tendenza all'innalzamento della vocali post-accentate nei proparossitoni in friulano (cfr. nota 31), possiamo ipotizzare che il processo armonico preceda l'innalzamento, come esemplificato in (11), dove in /ˈjevile/ *jevile* "alzala" il verbo ospite è /ˈjeve/ *jeve* (imperativo, 2a pers. sing. < lat. LEVA):

(11)		/ˈjeve+le/
		$[[\text{ˈj}\underset{\text{̣}}{\text{e}}\text{v}\underset{\text{̣}}{\text{ɛ}}]_{\text{PF}}\ \text{l}\underset{\text{̣}}{\text{ɛ}}]_{\text{PF}}$
	armonia	$[[\text{ˈj}\underset{\text{̣}}{\text{e}}\text{v}\underset{\text{̣}}{\text{e}}]_{\text{PF}}\ \text{l}\underset{\text{̣}}{\text{e}}]_{\text{PF}}$
	innalzamento	$[[\text{ˈj}\underset{\text{̣}}{\text{e}}\text{vi}]_{\text{PF}}\ \text{l}\underset{\text{̣}}{\text{e}}]_{\text{PF}}$

D'altra parte, l'ipotesi percettiva non va del tutto esclusa a priori. Come suggeriscono Jiménez & Lloret (c.d.s.) per il valenziano, si può supporre che i tratti armonici vengano estesi a un componente più prominente di quello rappresentato dalla sillaba accentata: il

[27] Si vedano, circa la possibilità di postulare un piede ternario, Bafile (1994, 1996), Loporcaro (2002). Sebbene sembrino non esistere, in friulano, esempi di proparossitoni con /e, o/ in sillaba accentata, i casi in (8) sono prove sufficienti per poter difendere la plausibilità del piede ternario.

[28] Per una sintesi della questione, cfr. anche Jiménez & Lloret (c.d.s.).

[29] In riferimento, nello specifico, all'esigua distanza che separa, in tutti i casi considerati, le vocali che inducono l'assimilazione e le vocali che ne assimilano i tratti, distanza inversamente proporzionale allo sforzo richiesto per mantenere il contrasto.

piede che ospita la vocale accentata, che induce il processo armonico. Quest'ipotesi presenta il vantaggio di spiegare in modo forse più agile i casi in (8), dal momento che evita la necessità di far appello a restrizioni circa la contiguità dei segmenti coinvolti; come già detto, infatti, il beneficio percettivo non comporta necessariamente un beneficio articolatorio, cosicché il dominio armonico può non essere omogeneo. Più verosimile, viste le prove addotte, e considerata la loro concilabilità, è propendere per la doppia interpretazione, articolatoria e percettiva nel contempo: il processo armonico verrebbe messo in atto, da una parte, per migliorare la percettibilità dei tratti diffusi, dall'altra, per render più economico il gesto coarticolatorio nella realizzazione della sequenza degli elementi compresi nel dominio di detto processo.

8. INNALZAMENTO E RIDUZIONE A SCHWA IN FRIULANO[30]

Uno sguardo alla panoramica data in (1) permette di ricostruire i vari percorsi alternativi seguiti da -A latina in area friulana e di metter in luce i fattori che possono aver determinato sia l'orientamento del percorso stesso (verso la serie anteriore o posteriore), sia i diversi gradi d'innalzamento.

Per questi ultimi, sembra esser decisivo il contesto V + /s/ (coincidente, come già detto, col morfema del plurale o della 2a pers. sing. dei verbi). Come si può notare, in tutti i casi in cui -A è stata interessata da qualche tipo d'alterazione, la variante più alta compare proprio davanti alla C sibilante, mentre il contrario non si verifica in nessun caso. Emblematica, a tal proposito, l'innovazione che ha interessato il friulano centrale di pianura (propria, inizialmente, dell'udinese), che presenta -/is/ (attestata sin dai più antichi documenti scritti), il che suggerisce una tendenza all'innalzamento che dovette manifestarsi molto precocemente (probabilmente attraverso realizzazioni molto chiuse di /e/).[31]

---

[30] I dati discussi in questa sezione si riferiscono al solo soggetto maschile e costituiscono un primo e parziale anticipo dei risultati d'una ricerca tuttora in corso, condotta su una base più ampia d'informanti. Va tuttavia detto che si tratta d'un processo, quello della riduzione a *schwa*, che trova applicazione in tutti gl'informanti, come abbiamo potuto appurare per mezzo d'un'attenta analisi uditiva, e nel friulano centro-orientale in genere. Seppur nella provvisorietà dei risultati, il presente paragrafo intende evidenziare l'esistenza d'importanti condizionamenti di vario tipo (contestuali e fonosintattici, innanzitutto) che influiscono sulle concrete realizzazioni dello *schwa* in friulano, mettendo in luce le caratteristiche variabili del processo osservato.

[31] Benincà & Vanelli (1978) mettono in relazione l'innalzamento di /a(s)/ finale con la generale tendenza all'innalzamento della vocali postaccentate nei proparossitoni in friulano (tranne davanti a /r/, che ha l'effetto opposto): /ˈwardʒ͡ine, ˈsabide, ˈkantilu, ˈkɔtule/ "aratro, sabato, cantalo, gonna" (lat. rispettivam. ORG/a/NU, SABB/a/TA, CANT/a/+LA, franc. *cotte* + suff. dim. romanzo -/o/*la*; le vocali in trascrizione fonologica sono quelle che vengono innalzate) *vs.* /ˈlɛtare/ "lettera" (lat. LITT/e/RA). Nel friulano centrale, la regola si applica a qualunque /a/ postaccentata, che viene innalzata di un grado (→ /e/) se in posizione finale (libera); si ha quindi un ulteriore innalzamento (→ /i/), che coinvolge, allo stesso tempo, /a/ innalzata e le vocali dello stesso grado d'altezza (/e, o/), davanti a qualunque C: /ˈmate/ "matta", /ˈmatis/ "matte" = /ˈwardʒ͡ine/. In altre varietà, invece (per esempio quelle che hanno /-as, -es/), l'innalzamento fino a /i/ riguarda solamente le vocali postaccentate davanti a C diverse da /s/: /ˈmatas, -es/ 'matte' ≠ /ˈwardʒ͡ine/. Dobbiamo però constatare una

Il contesto /Vs/ è stato dunque il primo ad aver intercettato il mutamento, in pressoché tutte le varietà considerate, anticipando d'almeno una tappa il processo d'innalzamento di -A, facendosi così promotore del processo stesso.

Oltre all'innalzamento, /s/ sembra favorire la centralizzazione della vocale precedente. I nostri dati mostrano che l'effetto-centralizzazione risulta considerevolmente amplificato in posizione non-prepausale, dove -/e/ è interessata da un processo 'macroscopico' (Romito *et al.*, 1997) di centralizzazione, che la riduce a [ə] (F1: 362/387 Hz, F2: 1481/1474 Hz),[32] come risulta evidente dalla Fig. 3 e dalla Tab. 3 (relative alla posizione dopo C dentali).[33]

---

certa asimmetria tra il paradigma nominale e quello verbale: l'equivalenza sing. femm. : 3ª pers. sing. pres. indic. = plur. femm. : 2ª pers. sing. pres. indic. non è rispettata da tutte le varietà; infatti, se lo è in friulano centrale, dove *mate : (al) cjante = matis : (tu) cjantis*, non lo è, per es., nelle varietà della fascia sudorientale del basso Tagliamento che hanno *mate = (al) cjante* ma *mates ≠ (tu) cjantis* (Sedegliano) o addirittura *mata = (al) cjanta* ma *mates ≠ (tu) cjantis* (Dignano). È ipotizzabile che l'esito *-is* che caratterizza la seconda persona singolare dei verbi della prima coniugazione, poi generalizzata, per analogia, anche per le altre coniugazioni (innovazione diffusa dall'udinese), sia stato favorito dall'esistenza, nello stesso paradigma, di terminazioni analoghe (*-i* + C#), nella fattispecie, quella della terza persona plur. *-in*, dove /i/ è il normale esito di /a/ postaccentata in un proparossitono, risultato dell'applicazione della regola vista: *a cjantin* 'cantano' < CANTANT, se consideriamo -NT eterosillabico rispetto alla vocale del tema (cfr. it. *cantano*, con *-o* d'appoggio). Questo porta a ritenere che, in friulano centrale, l'innalzamento categorico a /i/ in posizione postaccentata preconsonantica sia stato un processo morfologicamente condizionato, che avrebbe interessato in un primo momento i soli proparossitoni, per poi attivarsi anche davanti a /s/ morfema verbale (favorito da livellamento analogico), e imporsi, infine, anche davanti a /s/ morfema nominale.

[32] Per un termine di confronto, si considerino i valori medi calcolati per lo *schwa* da Schwartz *et al.* (1997): F1: 414 Hz, F2: 1516 Hz. I valori riportati riflettono, in realtà, il condizionamento esercitato dalle sole vocali medie anteriori accentate. Il limite è dovuto alla conformazione del corpus raccolto per questa fase della nostra indagine, che prevedeva la registrazione d'esecuzioni il più possibile spontanee d'enunciati, solo parzialmente guidate da un questionario, il che ci ha permesso di raccogliere un numero rappresentativo di stimoli vocalici per la maggior parte, ma non per tutti i contesti. Rimandiamo a uno studio futuro il completamento dell'indagine. Avvertiamo comunque che una prima valutazione sommaria, che tiene conto anche di tutti i contesti esclusi, ci consente di dire che il valore medio non subisce modificazioni sostanziali per quanto riguarda F1, mentre risulta relativamente inferiore per F2 (sintomo di posteriorizzazione lungo l'asse anteroposteriore, effetto, naturalmente, dell'inclusione dei contesti con vocale posteriore in sillaba accentata). Quel che in questa sede importa rilevare, tuttavia, non sono valori assoluti, bensì la loro importanza relativa all'interno d'un processo (quello, appunto, della riduzione a *schwa*), osservato dal punto di vista delle vocali medie anteriori (/ˈe, ˈɛ/.../e/, dove /ˈe/.../e/ rappresenta qualunque contesto armonico e /ˈɛ/.../e/ qualunque contesto non-armonico).

[33] Sebbene la differenza tra i valori di F1, nel confronto tra le realizzazioni in contesto armonico *vs* non-armonico, risulti ancora statisticamente significativa al *t* test ($p < 0,05$), percettivamente essa non pare sostanziale. Solo un'indagine estesa a ulteriori soggetti potrà, naturalmente, livellare le idiosincrasie e far decidere se sia giustificato utilizzare una sola

È importante segnalare che, nel soggetto maschile, la tendenza alla neutralizzazione degli effetti del condizionamento coarticolatorio, per -/es/ non-prepausale, è particolarmente evidente, sulla dimensione di F1, dopo C dentali e sibilanti (per queste ultime, l'effetto è già apprezzabile in posizione prepausale; v. Fig. 4),[34] mentre tali effetti tendono a persistere, per esempio, quando la vocale è preceduta da C vibrante, dove -/es/ dopo /ˈe/ è [ə̝] (F1: 390 Hz, F2: 1479 Hz) e dopo /ˈɛ/ è [ə̞ ~ ɜ̞] (F1: 480 Hz, F2: 1465 Hz). Come s'evince dalla Fig. 3, mentre le vocali accentate /ˈe, ˈɛ/ si mantengono ben distinte, conservando una collocazione sostanzialmente periferica, al mutare delle condizioni fonosintattiche (prepausale: FIN *vs.* non-prepausale: INT),[35] le non-accentate finali seguite da /s/, ancora ben differenziate in FIN, non risentono (quasi) più del condizionamento coarticolatorio a distanza, tendendo a neutralizzarsi in *schwa* (al contrario, se non seguite da /s/, continuano a mantenersi sufficientemente differenziate anche in INT, nonostante l'apprezzabile grado di centralizzazione). Si tratta, dunque, d'un fenomeno che va al di là dei normali processi di *undershoot* formantico, diafasicamente motivati (Romito *et al.*, 1997; Savy & Cutugno, 1997; Savy *et al.*, 2005). È inoltre notevole che la riduzione a *schwa* si verifichi anche a una velocità d'eloquio che possiamo definire 'normale' (o addirittura 'lenta'), che non risente dunque degli effetti imputabili a un'eccessiva ipoarticolazione.

---

etichetta fonetica o più etichette, nel contesto considerato. I dati rilevanti che emergono, ai fini della comprensione del processo sono, in ultima analisi: a) la più decisa tendenza alla neutralizzazione dei meccanismi armonici in contesto dentale rispetto ad altri contesti (nella fattispecie, quello vibrante), b) la relativa variabilità riscontrabile nella realizzazione di *schwa*, in dipendenza da fattori contestuali. Le differenze sono senz'altro significative e percettivamente rilevanti in tutti gli altri casi, cioè tra gli elementi di ciascuna coppia (-e/-e; -es/-es) definita in base al contesto fonosintattico. Per quanto riguarda F2, le differenze tra i valori medi in contesto armonico *vs* non-armonico risultano significative solo in posizione prepausale.

[34] La riduzione a *schwa* avviene dunque per effetto d'un progressivo e proporzionale arretramento, lungo la dimensione di F2, di tutte le categorie definite in base ai parametri visti (-/e/ FIN > -/es/ FIN > -/e/ INT > -/es/ INT), d'un concomitante innalzamento lungo l'asse verticale e d'un progressivo accorciamento delle distanze, lungo lo stesso asse, tra le realizzazioni di -/e/ nei contesti armonici e nei contesti non-armonici.

[35] Un analogo condizionamento contestuale dell'armonia vocalica è documentato per la varietà di Claro da Delucchi (2008: 268-291); cfr. nota 6.

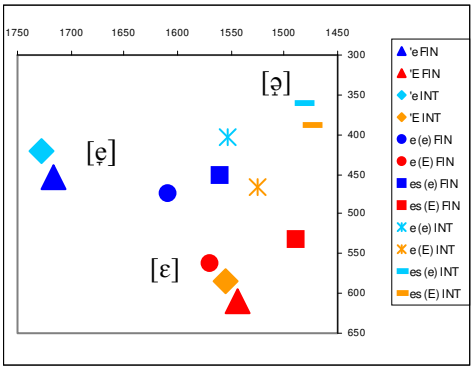


Figura 3: -/e(s)/ preceduta da C dentale, dopo V media anteriore (semichiusa *vs.* semiaperta, tra parentesi tonde nella figura) in friulano (valori medi; Tab. 3 per dettagli): effetti coarticolatori in posizione finale prepausale (FIN) e all'interno di frase (INT). Simboli SAMPA

	F1 media	F1 min	F1max	F1 ds	F2 media	F2 min	F2 max	F2 ds
('e)...e FIN	474	434	518	18	1609	1527	1716	44
('ɛ)...e FIN	563	536	619	21	1570	1519	1622	25
('e)...es FIN	450	436	476	11	1561	1505	1609	30
('ɛ)...es FIN	533	501	551	14	1489	1429	1533	28
('e)...e INT	403	333	454	46	1554	1471	1628	63
('ɛ)...e INT	467	399	513	38	1526	1454	1651	50
('e)...es INT	362	323	395	24	1481	1422	1540	35
('ɛ)...es INT	387	361	411	19	1474	1438	1533	28

Tabella 3: Valori medi e indici statistici (valore minimo, massimo e deviazione standard) di F1 e F2, relativi a -/e/ e -/es/ dopo C dentale (contesti armonici *vs.* non-armonici, dopo vocali medie anteriori) in posizione prepausale (FIN) *vs.* non-prepausale (INT)

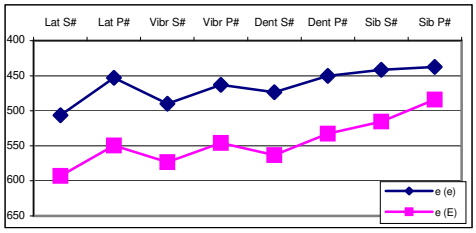


Figura 4: Condizionamento della C precedente sui valori di F1 relativi a -/e/ (S) e -/es/ (P) in posizione finale (soggetto maschile), dopo V media anteriore (semichiusa *vs.* semiaperta). Simboli SAMPA

Il quadro sin qui delineato per (1j) fornisce elementi per spiegare (1g), di cui potrebbe rappresentare (almeno per /Vs/) lo stadio immediatamente precedente. La situazione schematizzata in (1g), nel quadro delle varietà friulane dato in (1), corrisponde certamente allo sviluppo più innovativo. Le varietà indicate (i dati si riferiscono a un soggetto maschile anziano di Treppo C.) si trovano in un'area (Carnia orientale) che, grosso modo, ha quasi compiuto (se si eccettuano punti conservativi) l'innovazione *-a(s) → -e(s)*. È ipotizzabile che l'esito /əs/ provenga da un precedente /es/, e che la riduzione a *schwa*, dapprima fonosintatticamente condizionata (come lo è attualmente nella fascia sudorientale del basso Tagliamento, dove è ancora all'opera una regola allofonica), si sia poi generalizzata anche alla posizione prepausale, divenendo categorica e fonologizzandosi come /ə/. D'altra parte, non si può escludere l'ipotesi d'un percorso autonomo, che non presupponga necessariamente /es/,[36] e che /ə(s)/ provenga direttamente da un precedente /a(s)/, mediante progressivo innalzamento. Anche in questo caso, rimane comunque valida l'ipotesi che lo *schwa* si sia fonologizzato in seguito a una fase d'applicazione d'una regola allofonica, che prevedeva varianti centralizzate all'interno di frase e [a(s)] in posizione finale assoluta.

RINGRAZIAMENTI

Colgo l'occasione per ringraziare Maria Chiara Felloni per i materiali di prima mano, relativi al parmigiano *oltretorrentino*, gentilmente raccolti sul campo grazie a un'indagine da lei stessa condotta. Il mio ringrazionamento va anche a tutti gl'informanti friulani, catalani e valenziani che pazientemente hanno accettato di sottoporsi all'indagine, in particolare a Mauro, Laura ed Eugenio, di Sedegliano (UD), ad Anna, di Camino al Tagliamento (UD), ai sigg. Elena, Ugo e Giovanni, di Clauzetto (PN), ai miei informanti valenziani, Alicia, Neus e Ximo, a Paolo Roseano e Josefina Carrera-Sabaté, del Laboratorio di Fonetica dell'Università di Barcellona, a Luciano Canepari (Università di Venezia) e a Daniele Vitali. Si ringraziano infine Jesús Jiménez (Università di Valenza) e Maria-Rosa Lloret (Università di Barcellona) per i materiali fornitimi, e Giovanna Marotta (Università di Pisa), Michele Loporcaro (Università di Zurigo) e Mirko Grimaldi (Università di Lecce) per i preziosi suggerimenti. È dell'autore la responsabilità per eventuali errori, inesattezze o inadeguatezze presenti nel lavoro.

[36] -/e(s)/ per -/a(s)/ potrebbe esser infatti frutto d'un'innovazione seriore nell'area considerata, paracadutata dai modelli centrali pianigiani di maggior prestigio (ma, come detto, non accolta uniformemente in tutto il territorio), e prodottasi successivamente al compimento del processo -/a(s)/ > -/ə(s)/.

9. BIBLIOGRAFIA

Albano Leoni, F. & Maturi, P. (1998[2]), *Manuale di fonetica*, Roma: Carocci.

Ascoli, G. I. (1873), Saggi ladini, *Archivio Glottologico Italiano*, 1.

Avalle D'Arco, S. (2002), *La doppia verità. Fenomenologia ecdotica e lingua letteraria del medioevo romanzo*, Firenze: Edizioni del Galluzzo, 249-298.

Bafile, L. (1994), La riassegnazione postlessicale dell'accento in napoletano, *Quaderni del Dipartimento di Linguistica dell'Università di Firenze*, 5, 1-23.

Bafile, L. (1996), Sulla rappresentazione delle strutture metriche ternarie, *Quaderni del Dipartimento di Linguistica dell'Università di Firenze*, 7, 2-24.

Beckman, J.N. (1998), *Positional Faithfulness*, Tesi di dottorato, University of Massachusetts, Amherst (http://roa.rutgers.edu/).

Benincà, P. (1989), Friaulisch: interne Sprachgeschichte I. Grammatik. Evoluzione della grammatica, in *Lexikon der Romanistischen Linguistik* (G. Holtus *et al.*, editors), Tübingen: Niemeyer, III, 563-585.

Benincà, P. & Vanelli, L. (1975), Morfologia del verbo friulano: il presente indicativo, *Lingua e contesto*, 1, 1-62.

Benincà, P. & Vanelli, L. (1978), Il plurale friulano, contributo allo studio del plurale romanzo, *Revue de Linguistique Romane*, 42, 241-292.

Bertinetto, P.M. (1988), Reflections on the dichotomy 'stress' vs. 'syllable timing', *Quaderni del Laboratorio di Linguistica*, 2, 59-84.

Biondelli, B. (1853), *Saggio sui dialetti gallo-italici*, Milano.

Bocchialini, J. (1942), Come si pronuncia e si scrive il dialetto parmigiano, *I Quaderni de 'La Giovane Montagna'*, 97, 3-17.

Booij, G. (1996), Cliticization as prosodic integration: the case of Dutch, *The Linguistic Review*, 13, 219-242.

Calabrese, A. (1989), Phonological Variations, in *Dialect Variation and the Theory of Grammar. Proceedings of the Glow Workshop in Venice*, Venezia, 2 aprile 1987, 9-39.

Calamai, S. (2002), Vocali atone e toniche a Pisa, in *La fonetica acustica come strumento di analisi della variazione linguistica in Italia* (A. Regnicoli, editor), Atti delle XII Giornate di Studio del Gruppo di Fonetica Sperimentale, Macerata, 13-15 dicembre 2001, 39-46.

Canepari, L. (1999[2]), *Manuale di Pronuncia Italiana*, Bologna: Zanichelli.

Canepari, L. (2006[3]), *Manuale di fonetica*, München: Lincom.

Canepari, L. & Vitali, D. (1995), Pronuncia e grafia del bolognese, *Rivista Italiana di Dialettologia*, 19, 119-164.

Castellani, R. (1980), *Il friulano occidentale. Lineamenti storico-linguistici delle componenti dialettali*, Udine: Del Bianco.

Cole, J. & Kisseberth, C. (1994), An Optimal Domains Theory of Vowel Harmony, *Studies in the Linguistics Sciences*, 34, 101-114.

Delucchi, R. (2008), *Sui fenomeni di armonia vocalica nei dialetti della Svizzera italiana*, Tesi di Master, Università di Zurigo, Svizzera.

Dressler, W. U. (1985), *Morphonology: the dynamics of derivation*, Ann Arbor: Karoma.

Fant, G. (1960), *Acoustic Theory of Speech Production*, The Hague: Mouton.

Farnetani, E. (1997), Coarticulation and connected speech processes, in *The Handbook of Phonetic Sciences* (W. J. Hardcastle & J. Laver, editors), Oxford: Blackwell, 371-404.

Ferrero, F. (1979, a cura di), *L'identificazione della persona per mezzo della voce*, Roma: ESA.

Francescato, G. (1966), *Dialettologia friulana*, Udine / Tolmezzo: Società Filologica Friulana.

Frau, G. (1984), *Friuli*, Pisa: Pacini.

Gósy, M. (2004), The manifold function of schwa, *Grazer Linguistische Studien*, 62, 15-26.

Grimaldi, M. (2003), Quanto l'idea dei tratti distintivi è ancora un punto di raccordo fra fonetica e fonologia?, in *La coarticolazione* (G. Marotta & N. Nocchi, editors), Atti delle XIII Giornate di Studio del Gruppo di Fonetica Sperimentale, Pisa, 28-30 novembre 2002, Pisa: Edizioni ETS, 243-253.

Gsell, O. (1996), Chronologie frühromanischer Sprachwandel, in *Lexikon der Romanistischen Linguistik* (G. Holtus *et al.*, editors), Tübingen: Niemeyer, vol. II, 557-584.

Heilmann, L. (1961), Strutturalismo e storia nel dominio linguistico italiano: il vocalismo di una parlata tipica pavese, *Quaderni dell'Istituto di Dialettologia* (Univ. di Bologna), 6, 45-57.

Jiménez, J. (1998), Valencian Vowel Harmony, *Rivista di Linguistica*, 10, 137-161.

Jiménez, J. (2001), L'harmonia vocàlica en valencià, in *Actes del Novè Col·loqui d'Estudis Catalans a Nord-Amèrica*, Barcelona, Spagna, 1998, Barcelona: Publicacions de l'Abadia de Montserrat, 217-244.

Jiménez, J. (2002), Altres fenòmens vocàlics en el mot, in *Gramàtica del català contemporani* (J. Solà *et al.*, editors), Barcelona: Empúries, vol. 1, 171-194.

Jiménez, J. & Lloret, M.-R. (c.d.s.), Entre la articulación y la percepción: armonías vocálicas en la península Ibérica, in *Actes du XXV CILPR*, Innsbruck, Austria, 3-8 settembre 2007.

Ladefoged, P. & Maddieson, I. (1996), *The Sounds of the World's Languages*, Oxford: Blackwell.

Lafont, R. (1991), Okzitanisch: Interne Sprachgeschichte I. Grammatik. Histoire interne de la langue I. Grammaire, in *Lexikon der Romanistischen Linguistik* (G. Holtus *et al.*, editors), Tübingen: Niemeyer, vol. V.2, 1-17.

Lausberg, H. (1976[2]), *Linguistica romanza*, vol 1: *Fonetica*, Milano: Feltrinelli.

Loporcaro, M. (1999), Teoria fonologica e ricerca empirica sull'italiano ed i suoi dialetti, in *Fonologia e morfologia dell'italiano e dei dialetti d'Italia*, Atti del XXXI Congresso della SLI, Padova, 25 settembre 1997 (P. Benincà *et al.*, editors), Roma: Bulzoni, 117-151.

Loporcaro, M. (2000), Stress stability under cliticization and the prosodic status of Romance clitics, in *Phonological theory and the dialects of Italy* (L. Repetti, editor), Amsterdam / Philadelphia: John Benjamins, 137-168.

Loporcaro, M. (2002), Unveiling a masked change: behind vowel harmony in the dialect of Claro, in *Sounds and Systems. Studies in Structure and Change. A Festschrift for Theo Vennemann* (D. Restle & D. Zaefferer, editors), Berlin: Mouton de Gruyter, 75-90.

Loporcaro, M. (2005-06), I dialetti dell'Appennino tosco-emiliano e il destino delle atone finali nel(l'italo-)romanzo settentrionale, *L'Italia Dialettale*, 66-67, 69-122.

Loporcaro, M., Delucchi, R., Nocchi, N., Paciaroni, T. & Schmid, S. (2007), Schwa finali sull'Appennino emiliano: il vocalismo del dialetto di Piandelagotti, in *Scienze vocali e del linguaggio. Metodologie di valutazione e risorse linguistiche* (V. Giordani, V. Bruseghini & P. Cosi, editors), Atti del 3° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, 29 novembre – 1° dicembre 2006, Povo (Trento), Torriana (RN): EDK Editore, 57-76.

Manuel, S. Y. & Krakow, R. A. (1984), Universal and language particular aspects of vowel-to-vowel coarticulation, *Haskins Laboratories Status Report on Speech Research*, SR-77/78, 69-78.

Marinelli, G. & Gortani, M. (1924-25), *Guida della Carnia e del Canal del Ferro*, Tolmezzo: Stab. Tip. Carnia.

Marotta, G. & Savoia, L.M. (1991), Diffusione vocalica in un dialetto calabrese. Alcuni parametri fonologici, in *L'interfaccia tra fonologia e fonetica* (E. Magno-Caldognetto & P. Benincà, editors), Atti del Convegno, Padova, 15 dicembre 1989, Padova: Unipress, 19-42.

Miotti, R. (2002), Friulian, *Journal of the International Phonetic Association*, 32, 237-247.

Miotti, R. (2007), Le varietà di Dignano, Flaibano e Sedegliano nel contesto dei dialetti friulani. Aspetti fonologici, in *Ladine loqui. IV Colloquium Retoromanistich*, San Daniele, 26-27 agosto 2005 (F. Vicario, editor), Udine: Società Filologica Friulana, 71-117.

Nespor, M. (1993), *Fonologia*, Bologna: il Mulino.

Nespor, M. & Vogel, I. (1986), *Prosodic phonology*, Dordrecht: Foris.

Nievo, I. (2006), *Confessioni d'un italiano*, Torino: UTET (ed. a cura di L. M. Marchetti).

Ohala, J.J. (1989), Sound change is drawn from a pool of synchronic variation, in *Language Change, Contributions to the Study of its Causes* (L.E. Breivik & E.H. Jahr, editors), Berlin: Mouton de Gruyter, 173-198.

Ohala, J. J. (1993), Coarticulation and phonology, *Language and Speech*, 36, 155-170.

Paradis, C. & Prunet, J.-F. (1991, editor), *The Special Status of Coronals: Internal and External Evidence, Phonetics and Phonology*, vol. 2, San Diego: Academic Press.

Pellegrini, R. (1994), Friuli, in *Storia della lingua italiana* (L. Serianni & P. Trifone, editors), Torino: Einaudi, vol. III, 240-260.

Pellicer, J. & Giner, R. (1997), *Gramática de uso del valenciano*, Valencia: mil999.

Pulleyblank, D. (2002), Harmony drivers: no disagreement allowed, in *Proceedings of the Twenty-eighth Annual Meeting of the Berkeley Linguistics Society*, Berkeley, California (J. Larson & M. Paster, editors), Berkeley Linguistics Society, 249-267.

Recasens, D. (1991), *Fonètica descriptiva del català*, Barcelona: Institut d'Estudis Catalans.

Rohlfs, G. (1966[2]), *Grammatica storica della lingua italiana e dei suoi dialetti*, I: *Fonetica*, Torino: Einaudi.

Romito, L. Turano, T., Loporcaro, M. & Mendicino, A. (1997), Micro e macrofenomeni di centralizzazione nella variazione diafasica: rilevanza dei dati fonetico-acustici per il quadro dialettologico calabrese, in *Fonetica e fonologia degli stili dell'italiano parlato*, Atti delle VII Giornate di Studio del GFS, Napoli, 14-15 dicembre 1996 (F. Cutugno, editor), 157-174.

Salvioni, C. (1892-94), L'influenza della tonica nella determinazione dell'atona finale in qualche parlata della valle del Ticino, *Archivio Glottologico Italiano*, 13, 355-360.

Sánchez Miret, F. (1999), Assimilazione a distanza fra vocali nei dialetti italiani: fonetica e spiegazione del cambiamento, in *Fonologia e morfologia dell'italiano e dei dialetti d'Italia* (P. Benincà, A. Mioni & L. Vanelli, editors), Atti del 31° Congresso della Società di Linguistica Italiana, Padova, 25-27 settembre 1997, Roma: Bulzoni, 269-290.

Savy, R. & Cutugno, F. (1997), Ipoarticolazione, riduzione vocalica, centralizzazione: come interagiscono nella variazione diafasica?, in *Fonetica e fonologia degli stili dell'italiano parlato* (F. Cutugno, editor), Atti dell VII Giornate di Studio del Gruppo di Fonetica Sperimentale, Napoli, 14-15 dicembre 1996, 177-194.

Savy, R., Lo Prejato & Clemente, G. (2005), Per una caratterizzazione e una misura della riduzione vocalica in italiano, in *Misura dei parametri. Aspetti tecnologici e implicazioni nei modelli linguistici* (P. Cosi, editor), Atti del 1° Convegno dell'Associazione Italiana di Scienze della Voce, Padova, 2-4 dicembre 2004, Torriana (RN): EDK Editore, 135-160.

Schwartz, J.-L., Boë, L.-J., Vallée, N. & Abry, C. (1997), The Dispersion-Focalization Theory of vowel systems, *Journal of Phonetics*, 25, 255-286.

Sganzini, S. (1924-26), Fonetica dei dialetti della Val Leventina, *L'Italia Dialettale*, 1, 190-212; 2, 100-155.

Sorianello, P. (2002), Il vocalismo tonico senese: un'indagine sperimentale, in *La fonetica acustica come strumento di analisi della variazione linguistica in Italia* (A. Regnicoli, editor), Atti delle XII Giornate di Studio del Gruppo di Fonetica Sperimentale, Macerata, 13-15 dicembre 2001, 47-52.

Tagliavini, C. (1972[6]), *Le origini delle lingue neolatine*, Bologna: Patron.

Van Bergem, D. (1994), A model of coarticulatory effects on the schwa, *Speech Communication*, 14, 143-162.

Van Der Hulst, H. & Van De Weijer, J. (1996), Vowel Harmony, in *The Handbook of Phonological Theory* (J.A. Goldsmith, editor), Oxford: Blackwell, 495-534.

Vanelli, L. (1998), Le vocali lunghe del friulano, *Quaderni della Grammatica Friulana di Riferimento*, 1, 69-108.

Veny, J. (1998[12]), *Els parlars catalans (síntesi de dialectologia)*, Palma de Mallorca: Moll.

Vicario, F. (1999), *Il quaderno dell'Ospedale di Santa Maria Maddalena*, Udine: Biblioteca Civica 'V. Joppi'.

Vicario, F. (2001), *Carte friulane del Quattrocento dall'archivio di San Cristoforo di Udine*, Udine: Società Filologica Friulana.

Walker, R. (2005), Weak Triggers in Vowel Harmony, *Natural Language and Linguistic Theory*, 23, 917-989 (http://roa.rutgers.edu/).

Walker, R. (2006), Long-distance Metaphony: A Generalized Licensing Proposal, Lavoro presentato al *PhonologyFest Workshop*, Indiana University, Bloomington (http://www-rcf.usc.edu/~rwalker/pubs.html).

Zamboni, A. (1990), Per una riconsiderazione generale del vocalismo cisalpino: l'abbassamento di /e/ neolatino in posizione, in *Scritti in onore di Lucio Croatto*, Padova: Centro di studio per le ricerche di fonetica del CNR, 287-296.

Zamboni, A. (1995), Per una ridefinizione del tipo alto-italiano o cisalpino, in *Italia settentrionale: crocevia di idiomi romanzi* (E. Banfi *et al.*, editors), Atti del convegno internazionale di studi, Trento, 21-23 ottobre 1993, Tübingen: Max Niemeyer Verlag, 57-67.

COARTICOLAZIONE E MUTAMENTO. UNA RICERCA SUL VOCALISMO ATONO FINALE NELL'ENTROTERRA MACERATESE

Tania Paciaroni
Università di Zurigo
paciaron@rom.uzh.ch

1. SOMMARIO

Come è noto, l'aspetto tipizzante nel vocalismo atono finale dell'area mediana è la distinzione di /-u/ ed /-o/ finali, che non sono confluite in /-o/ come in italiano standard. All'interno dell'area maceratese-fermana-camerte, fin dai primi decenni del secolo scorso, tuttavia, Erich Mengel (1936) segnalava, nelle varietà del litorale costiero, la sostituzione della /-u/ con /-o/, per effetto dell'influenza dello standard, di contro alla saldezza della distinzione nell'interno. L'attualità di questo quadro è a tutt'oggi confermata dalla letteratura linguistica (Balducci, 2000: 28-29). Ne è *exemplum*, tra gli altri, la parlata di Matelica, in provincia di Macerata, ove i parlanti riconoscono nel fatto di "avere più *-u*" il tratto caratteristico del proprio dialetto. Tuttavia, a dispetto di questa autovalutazione e delle descrizioni disponibili (cfr. Bricchi, 1984; Traballoni, 2002-2003), a chi ascolti il matelicese non sfuggirà la variazione timbrica delle /-u/ finali.

Per verificare se tale variabilità sia un tratto evolutivo recente, dettato da interferenza con l'italiano, o se costituisca invece una tappa intermedia verso una configurazione diversa, eventualmente di tipo già noto e presente nell'area (cfr. Paciaroni & Loporcaro, in stampa), si è proceduto alla raccolta di materiale in inchieste sul campo con tre informatori dialettofoni in tre stili enunciativi diversi: parole in isolamento, entro frase e in parlato (semi)spontaneo. Dal *corpus* sono state estratte le sequenze contenenti: i) /o/ tonica; ii) /u/ tonica; iii) /-o/ finale < -ŏ, -ō; iv) /-u/ finale dell'articolo, del clitico oggetto e del dimostrativo; iv) /-o/ finale dell'articolo, del clitico oggetto e del dimostrativo; v) /-u/ delle altre categorie lessicali con base etimologica -ŭ(M). Per ogni vocale, in tutti e tre gli stili di parlato, sono state misurate la frequenza delle prime due formanti (F1, F2), la durata del segmento, l'intensità massima. Complessivamente sono stati misurati 612 stimoli.

Dalle analisi è risultato che l'abbassamento di /-u/ che oggi si presenta a Matelica è un fenomeno di variazione soggetto a condizioni coarticolatorie, senza ripercussioni sul sistema (e perciò non accessibile alla coscienza metalinguistica del parlante), dipendente dalla tipologia di parlato. L'indagine ha anche mostrato che la variazione non è generalizzata, ma ristretta alle sole parole lessicali, mentre non riguarda le parole funzionali, ove l'opposizione /-u/ ≠ /-o/ svolge la funzione morfologica di marcamento di genere maschile ≠ neutro.

Il dialetto di Matelica sembra quindi rispecchiare una tappa intermedia del percorso diacronico che, a partire da una configurazione di tipo reatino, con realizzazione di /-u/ univoca ([-u]), ha portato all'insorgere di sistemi con armonia vocalica di tipo cervaròlo (Merlo, 1922), ampiamente documentati nell'area mediana, presenti nel maceratese, tra gli altri, a San Severino M., con /-u/ che si abbassa a [-o] dopo vocale tonica media, indipendentemente dal fattore diafasico.

2. INTRODUZIONE*

In questa comunicazione si analizza il vocalismo atono finale di una varietà dell'entroterra maceratese, a partire da differenti punti di vista, dialettologico, fonetico e geolinguistico.

Come è noto, l'aspetto tipizzante nel vocalismo atono finale dell'area mediana, a cui il maceratese appartiene, è la distinzione di /-u/ ed /-o/ finali, che non sono confluite in /-o/ come in italiano standard. Fin dai primi decenni del secolo scorso, tuttavia, Mengel (1936: 18ss.) segnalava la sostituzione, nelle varietà del litorale costiero, della /-u/ con /-o/, per effetto dell'influenza dello standard, di contro alla saldezza della distinzione nell'interno: "der -u Auslaut eine umso größere Vitalität zeigt, je höher und weiter man in das gebirgige Hinterland des nördlichen Picenums eindringt. Im Küstengebiet um Civitanova ist er so gut wie verschwunden und dort gilt er heute als das charakteristische Merkmal der parlata dei montanari". L'attualità di questo quadro è a tutt'oggi confermata dalla letteratura linguistica (v. Balducci, 2000: 28-29). Ne è *exemplum*, tra gli altri, la parlata di Matelica, in prov. di Macerata, ove i parlanti riconoscono nel fatto di "avere più *-u*" il tratto caratteristico del proprio dialetto (cfr. *infra*, §4.3). Tuttavia, a dispetto di questa autovalutazione, a chi ascolti il matelicese non sfuggirà la variazione timbrica delle /-u/ finali.

È lecito a questo punto chiedersi se la vitalità di /-u/ di cui Mengel scriveva non stia venendo meno anche nell'interno, vittima della 'regressione dialettale' comune alle varietà italo-romanze in genere, condivisa anche da Matelica (cfr. Gebhardt, 2007). Ma si può anche ipotizzare che sia in atto il mutamento verso una nuova distribuzione delle /-u/ ed /-o/ finali, che, come mostrato in Paciaroni & Loporcaro (in stampa), hanno configurazioni diverse nelle varietà odierne dell'area maceratese. O ancora, si può supporre che si realizzi a Matelica un fenomeno di variazione allomorfica fonetico-prosodica quale quella illustrata da Maturi & Schmid (1999, 2001, 2002 e 2003) nei dialetti campani.

Alla luce di queste possibili spiegazioni, il presente studio analizza le realizzazioni delle *-u* finali nel dialetto di Matelica, mettendone a fuoco in particolare i correlati acustici, nonché le ripercussioni sul sistema.

Il lavoro si apre con una rassegna sulla letteratura relativa all'area linguistica cui i dati fanno riferimento (§3), segue una messa a fuoco della parlata di Matelica (§4), ove la presenza/assenza dell'opposizione fra /-u/ e /-o/ finali viene verificata a partire dalle fonti scritte linguistiche (§4.2) e non linguistiche (§4.3) e facendo riferimento all'analisi percettiva (§4.4). La sezione successiva ha per oggetto una descrizione dettagliata del disegno sperimentale adottato nell'analisi acustica (§5); vengono poi descritti i risultati dell'analisi acustica per le parole funzionali (§6.1) e per le parole lessicali (§6.2). Per valutare meglio le implicazioni dei risultati ottenuti, con riferimento in particolare ai rapporti con gli altri dialetti della zona, viene inoltre svolto un confronto con altre varietà italo-romanze (§§6.1.2, 6.2).

* Una parte di questo lavoro è stata presentata al XXXII Convegno annuale della Società Italiana di Glottologia (S.I.G.), Verona, 25-27 ottobre 2007 (Paciaroni, in stampa).

3. IL VOCALISMO ATONO FINALE NEI DIALETTI MEDIANI

Nelle varietà mediane sono del tutto assenti tanto l'apocope quanto la centralizzazione dei timbri vocalici finali. Grazie al mantenimento della distinzione di -U e -O finali latine vi si ritrova, come noto, il tipo più conservativo tra quelli italo-romanzi:

(1)
```
-i        -u
   -e  -o
      -a
```

Per questo tratto, documentato fin dai primi testi in volgare, valgano, anche come contrappunto antico, le parole di Carlo Salvioni (1900: 7) nell'edizione del trecentesco *Pianto delle Marie*: "[l]a distinzione tra *-u* e *-o* (Meyer-Lübke, *It.Gr.* §109) è sempre osservata nel senso che allato a *-u* possa bensì comparire *-o*, ma mai non s'abbia *-u* per *-o*". Qui di séguito si esemplifica con forme matelicesi:

(2) a. *-u* < -Ŭ(M): ['fruttu] 'frutto', ['ʃordu] 'sciolto', ['ruʃʃu] 'rosso'
b. *-o* < -Ŏ, -Ō: ['d̥ɔrmo] 'dormo', [d̥ur'mi:mo] 'dormiamo',
['k̥o'rɛnno] 'correndo', ['io] 'io', ['k̥wanno] 'quando'

In tutta l'area mediana sulla distinzione /-u/ vs. /-o/ si è innestata, a partire dal sistema dell'articolo,[1] una distinzione morfologica di genere tra un maschile numerabile e un neutro non numerabile (cfr. almeno Rohlfs, 1949 [1968]: §419; Contini, 1961-1962; Parrino, 1967; Avolio,1996):

(3) a. [lu 'fjo:re] 'il fiore'
b. [lo 'sa:le] 'il sale'

In Paciaroni & Loporcaro (in stampa) si è mostrato che tale distinzione ha estensione diversa tra le varietà odierne dell'area maceratese; nel tipo prevalente, esemplificato in (4) dal maceratese urbano, essa ha raggiunto participio, aggettivo e nome:[2]

[1] Si aderisce qui all'ipotesi di Merlo (1906-1907), secondo cui *lu* < ĬLLŬM ≠ *lo* < *ĬLLOC o ĬLL'HOC rifatto su HOC. Per gli argomenti a sostegno di quest'ipotesi cfr. Paciaroni & Loporcaro (in stampa).

[2] Si tratta di una situazione per quest'area già ben nota grazie alla bella descrizione di Amerindo Camilli (1929) della parlata di Servigliano.

(4)		i. participi	ii. aggettivi	iii. nomi
-u m.	a.	[lu preˈʃuttu aˈdɛ ffiˈniːtu/*-o] ‘il prosciutto è finito’	[lu preˈʃuttu ˈkottu/*-o] ‘il prosciutto cotto’	[lu ˈfer(r)u] ‘il ferro (oggetto)’ (pl. [li ˈfer(r)i])
-o n.	b.	[lo ˈvi aˈdɛ ffiˈniːto/*-u] ‘il vino è finito’	[lo ˈvi kˈkotto/*-u] ‘il vino cotto’	[lo ˈfer(r)o] ‘il ferro (metallo)’

In altre varietà, la distinzione *-u* maschile ≠ *-o* neutro ricorre solo se etimologicamente originaria, ma la realizzazione di /-u/ < -ŭ(M) oscilla tra i gradi di apertura alto e medio-alto per l’agire dell’armonia vocalica. Così, tra gli altri, a San Severino M.,[3] ove si è prodotto un processo di armonia vocalica sensibile alle stesse condizioni individuate da Merlo (1922: 53) per la Cervara di Roma: «L’*-ǫ* (=cl. -ŭ) delle voci piane si règola secondo la qualità della tonica: è *-u* se la vocale della sillaba tonica è un *i*, un *a* o un *u* cervaroli; è *-o*, se nella tonica sono *ẹ*, (*e̲*), *ǫ*, (*o̲*)».[4] Di questa regolarità fonetica nel sistema sanseverinate rendono conto i dati illustrati in (5):

(5)		i. participi	ii. aggettivi	iii. nomi
-u/*-o* m.=n.	a.	[lu preˈʃuttu aˈdɛ ffiˈniːto/*-u] ‘il prosciutto è finito’	[lu preˈʃuttu ˈkotto/*-u] ‘il prosciutto cotto’	[lu ˈfer(r)o] ‘il ferro (oggetto)’ (pl. [li ˈfer(r)i])
	b.	[lo ˈvi aˈdɛ ffiˈniːtu/*-o] ‘il vino è finito’	[lo ˈvi kˈkotto/*-u] ‘il vino cotto’	[lo ˈfer(r)o] ‘il ferro (metallo)’

Siamo all’interno dell’Italia mediana metafonetica. Non stupisce, pertanto, l’ampia documentazione di un processo come questo che ha nella coarticolazione il suo sostrato materiale.[5]

4. MATELICA

4.1 Premessa

Il comune di Matelica, a 45 chilometri a ovest di Macerata e a 354 m.s.l.m., si sviluppa su un’altura al centro della valle del fiume Esino e conta poco più di 10,000 abitanti. Abitata da Piceni già dal IX secolo a.C., *Matilica* divenne Municipio romano nell’89 d.C. Nella suddivisione delle *regiones Augusteae* Matelica fu assegnata all’Umbria e solo dopo la caduta dei Ducati longobardi rientrò nei confini delle Marche attuali.

[3] I dati provengono dai *Materiali per il* Vocabolario del dialetto di San Severino Marche curato da Adriano Biondi e da inchieste svolte nell’estate 2007.

[4] Lo stesso processo avviene a Subiaco (Lindstrom, 1907), a Vallepietra (Merlo, 1930; Schirru, 2009) e in altri centri della valle dell’Aniene (Merlo, 1930).

[5] Per la metafonia cfr. almeno Maiden (1989, 1991); per i fenomeni di armonizzazione in area mediana cfr. Camilli (1929); Maiden (1988, 1995); Pucciarelli (2006); Schirru (2009).

4.2 Status quaestionis*: fonti linguistiche*

La realizzazione della distinzione tra /-u/ e /-o/ nella parlata di Matelica è fatto ben noto in letteratura. Nel suo elenco delle parlate presentanti l'opposizione *-u* ≠ *-o*, Merlo (1920: 260-261) annovera Matelica, adducendo le forme "*campu*, *lampu*, *sbagliu*, *altru*, *fattu*, di c. a *credo*, *voglio*, *dicenno*, *quanno*" (da Leopardi, 1887: 69).

La distinzione /-u/ vs. /-o/ veicola l'opposizione di genere maschile vs. neutro. Così Michela Traballoni (2001-2002: 46-47) descrive il 'neoneutro' matelicese:

> "il genere neutro si conserva quindi nell'articolo *lo* [lo], usato davanti a sostantivi che indicano concetti collettivi (*lo pilu* ['lo pilu], *lo cece* [lo 'ʧeʧe], *lo granu* [lo 'granu]), nomi di materia (*lo surfu* [lo 'surfu], *lo piummu* [lo 'pjummu], *lo carbó'* [lo kar'bo], nomi di sostanze liquide, solide e aeriformi (*lo vinu* [lo 'vinu], *lo lignu* [lo 'liŋŋu], *lo fume* [lo 'fume])"

La distinzione /-u/ vs. /-o/ si ritrova, dunque, nell'articolo (nonché nel clitico e nel dimostrativo ove è motivata etimologicamente), ma non si estende ai participi, agli aggettivi e ai nomi: in queste classi di parole l'uscita è sempre /-u/, sia per il genere maschile sia per il genere neutro, secondo condizioni etimologiche strutturalmente identiche a quelle note in letteratura per varietà come la reatina (v. Campanelli, 1896).

Nelle trascrizioni di etnotesti raccolti dalla studentessa romana Anita Lorenzotti (2000: 279) la /-u/ presenta, però, realizzazione variabile:

(6) "C: l at'tretsi 'kwanno s an'nava 'sui 'kampi […]
G: 'ɛra la perti'kara de 'leɲo o'pure 'kwella de 'fɛro […] lu vorda'rekkje
C: […] lu fatʃe'a:te lu 'sugu 'fintu"

Di fronte a questi dati, sono disponibili due spiegazioni. Una suppone un mutamento in atto in direzione del tipo maceratese, con opposizione di genere maschile /-u/ vs. neutro /-o/ anche ai sostantivi, l'altra, più semplice, e interna, un abbassamento di /-u/ a [-o] condizionato dalla vocale tonica media.

4.3 Informal literature*. Coscienza metalinguistica dei parlanti*

Della conservatività del loro vocalismo atono finale, i matelicesi hanno fiera consapevolezza. Lo testimoniano le parole di Amedeo Bricchi (1984: 7):

> "C'è anzitutto da dire che taluni, specialmente quelli dell'Italia settentrionale, hanno la sensazione che il nostro dialetto matelicese sia un linguaggio triviale e grossolano, sensazione che è alimentata dalle molte parole con la vocale u; invece questo è segno evidente di una accentuata permanenza dei resti del latino, che ha molte terminazioni in -u, -us, -um, -unt, -uc, -ur, ecc".

L'ispezione nei testi dialettali pubblicati conferma, attraverso l'uso grafico, la compattezza della categorizzazione come /-u/ nella coscienza dei parlanti:

(7) a. Boldrini (2006: 40): "Non fiateamo; io, benché di' volia / Centu cose, 'n sapio do' comincia'; Tu me fissai in un modu che paria / Che me arriasti l'annima a

fora’”

b. Baldini (2006: 48): “Io, un po’ scontroso, annài dendro l’acqua, rognichènno e un po’ biastimènno, perché, essènno scarzu, li sassi me piccàa. […] mangu dopo un metru ho troàta l’acqua arda. Io non lo sapìo e poi non so notà. Co’ l’acqua io non ce vo tandu d’accordu, perché me piace de più lo vinu”

4.4 Analisi percettiva

In realtà, ascoltando il parlato connesso, si percepiscono realizzazioni variabili di /-u/. Lo si esemplifica con il documento sonoro qui di séguito inserito, raccolto da chi scrive durante inchieste sul campo {audio 1}:

(8) NaC: perˈkᶣe na vˈɔlʈa s uˈsaːva ˈdenʈro le ˈkᶣaːse kee su le sofˈfitte tʃe vutˈtaa lo ˈɣraːnu ʃˈʃorʈo e pˈpoj tʃe metˈtiːa+ ˈpuːre le ˈmeːle pe fˈfalle matuˈra nno, non tʃe se metˈtiːa le ˈmeːle?

NC: si si, le ˈmeːle le me+, e ˈmeːla koˈtoɲɲe

GC: su, su lu ˈsakkʊ ðe lo ˈɣraːnu lo forˈmaddʒu a fˈfallo matuˈra mˈmɛʝʝo

NC: ˈvɛne ˈpju bˈboːno lo ˈkaːʃu ko la faˈriːna ɛra pˈpju bˈbʷona ˈðɔːᵽo nɔ, adˈdʒa ttʃ ˈɛːra tʃ ˈɛːra lo ˈkaʃʊ miʃˈkjaːʈ+

NaC: s ˈɛ ntsapoˈriːta

NaC: [lett.] ‘perché una volta si usava, dentro le case / che sulle soffitte ci si buttava il grano sciolto / e poi ci si mettevano pure le mele per farle maturare no / non ci si mettevano le mele?

NC: Sì sì, le mele, le mele cotogne

GC: su, nel sacco del grano, il formaggio per farlo maturare meglio

NC: viene più buono, il formaggio, con la farina, era più buono dopo, no / già c’era il formaggio mischiato

NaC: si è insaporita’

Di /-u/ in parlato (semi)spontaneo, dunque, l’analisi percettiva registra più di due varianti: [u ʊ o], la cui distribuzione è indipendente dalla posizione (interna o finale) nella frase e nel sintagma, non vincolata dal fattore di pausa che si realizza in presenza di confine sintattico. A questo punto abbiamo titolo per supporre che realizzazioni variabili di /-u/ esistano e siano determinate dal contesto segmentale. Qualora quest’ipotesi risultasse confermata, il dato, come visto *supra* (§3), non sarebbe nuovo per l’area, ove già altre varietà esibiscono realizzazioni diverse di /-u/ originaria, regolate dall’armonia vocalica.

5. L’INCHIESTA: MATERIALI E METODI

Il materiale utilizzato per il presente studio è frutto di inchieste condotte a più riprese dal 2006 al gennaio 2009 con sei informatori dialettofoni (4 maschi e 2 femmine), nati e residenti a Matelica (cfr. tabella 1):

Soggetto	Informatore	Sesso	Anno di nascita
1	GC	M	1930
2	NaC	M	1961
3	NC	M	1940
4	TT	M	1949
5	PS	F	1950
6	EP	F	1927

Tabella 1: Dati personali degli informatori

All'interno di questo corpus, ai fini dell'analisi che qui si presenta, sono state analizzate le vocali dei parlanti GC, NaC e NC in tre stili enunciativi diversi: parole in isolamento (PI), entro frase (PF) e in parlato (semi)spontaneo (PSp). Il metodo ha prodotto differenze stilistiche di natura quantitativa.

Le registrazioni sono state effettuate in abitazioni private. Per la segmentazione e l'etichettatura dell'intero corpus si è impiegato il software *Multi-Speech* 3700 (versione 2.5).

Per ogni vocale, in tutti e tre gli stili di parlato, sono stati misurati i parametri seguenti:

1) frequenza delle prime due formanti (F1, F2);
2) durata del segmento;
3) intensità massima.

La misurazione delle formanti è stata effettuata sull'intera durata del segmento vocalico mediante l'algoritmo LTA (*Long Term Average*), spettro medio a lungo termine che rappresenta la media di una serie di inviluppi spettrali calcolati con algoritmo di tipo FFT entro una porzione selezionata di segnale. Per la rappresentazione grafica dello spazio vocalico, i risultati acustici sono stati elaborati con Systat.

Dal *corpus* sono state segmentate ed etichettate le sequenze contenenti:

i) /o/ tonica;
ii) /u/ tonica;
iii) /-o/ finale;
iv) /-u/ finale dell'articolo, del clitico oggetto e del dimostrativo;
v) /-o/ finale dell'articolo, del clitico oggetto e del dimostrativo;
vi) /-u/ delle altre categorie lessicali in cui ritroviamo la base etimologica -Ŭ(M).

Per verificare l'ipotesi che la variabilità sia determinata da coarticolazione (come a San Severino M.), e non da livellamento analogico (come a Macerata) o da interferenza con l'italiano, si è proceduto alla disaggregazione delle realizzazioni di /-u/ secondo i) categoria lessicale; ii) genere; iii) qualità della vocale tonica della parola in cui /-u/ si trova.

I questionari delle parole in isolamento e delle parole in posizione interna di frase sono stati costruiti in modo da contenere le stesse parole bersaglio, differenti per categoria lessicale, per qualità della vocale tonica, e per contesto consonantico; entro parlato spontaneo, ove questo controllo non è possibile, si è badato a selezionare parole con caratteristiche (categoria lessicale, conformazione segmentale) analoghe. Sono state misurate 10 ricorrenze di ogni vocale per ciascun parlante in ciascuno dei tre stili; solo per /-o/ finale le ricorrenze analizzate sono inferiori, pari a 25 per il parlato in isolamento, 23 per il parlato entro frase, 22 per il parlato (semi)spontaneo. Complessivamente sono stati misurati 612 stimoli.

6. RISULTATI DELL'ANALISI ACUSTICA

Nelle tabelle 2-4 sono riportati i valori medi e la deviazione standard della prima e della seconda formante, della durata, dell'intensità riferite alle parole in isolamento (Tab. 2), alle parole entro frase (Tab. 3), alle parole in parlato (semi)spontaneo (Tab. 4).

La rappresentazione nello spazio acustico dei risultati è presentata in base alla categoria lessicale della parola bersaglio. In 6.1. si riportano i risultati relativi alle parole funzionali, in 6.2. quelli riferiti alle parole lessicali. I sottoparagrafi sono ordinati in base allo stile con cui è realizzata la parola bersaglio: prima vengono presentati i valori delle vocali all'interno di parole bersaglio pronunciate in isolamento, seguono i valori di vocali all'interno di parole entro frase, infine sono presentati i valori delle parole in parlato (semi)spontaneo.

/V/	[V]	Contesto	F1	F2	D	I
/ˈo/	[ˈo]		421 (17)	837 (46)	137 (34)	68 (1)
/ˈu/	[ˈu]		282 (12)	763 (55)	120 (6)	69 (1)
/-o/	[-o]		493 (34)	910 (19)	107 (8)	67 (4)
/-o/	[-o]	(articolo n.)	438 (11)	957 (26)	68 (15)	68 (8)
/-u/	[-u]	(articolo m.)	303 (27)	809 (45)	70 (9)	67 (3)
/-u/	[-o]	/ˈe ˈɛ ˈo ˈɔ/	311 (24)	782 (41)	104 (30)	68 (6)
	[-u]	/ˈu ˈa ˈV + C palatale/	287 (19)	786 (103)	105 (12)	68 (4)

Tabella 2: Valori medi e deviazione standard di /u/ e /o/ toniche e finali atone (PI)

/V/	[V]	Contesto	F1	F2	D	I
/ˈo/	[ˈo]		417 (13)	924 (31)	104 (7)	72 (7)
/ˈu/	[ˈu]		270 (10)	792 (22)	101 (2)	72 (6)
/-o/	[-o]		446 (31)	966 (66)	65 (15)	70 (1)
/-o/	[-o]	(articolo n.)	436 (26)	969 (28)	67 (6)	70 (7)
/-u/	[-u]	(articolo m.)	316 (51)	798 (76)	68 (7)	69 (8)
/-u/	[-o]	/ˈe ˈɛ ˈo ˈɔ/	392 (12)	928 (23)	71 (1)	66 (9)
	[-u]	/ˈu ˈa ˈV + C palatale/	317 (15)	814 (105)	69 (4)	69 (6)

Tabella 3: Valori medi e deviazione standard di /u/ e /o/ toniche e finali atone (PF)

/V/	[V]	Contesto	F1	F2	D	I
/ˈo/	[ˈo]		425 (46)	929 (52)	103 (13)	73 (6)
/ˈu/	[ˈu]		319 (48)	855 (69)	81 (6)	69 (3)
/-o/	[-o]		439 (35)	1050 (42)	69 (16)	71 (5)
/-o/	[-o]	(articolo n.)	436 (17)	981 (31)	68 (12)	69 (1)
/-u/	[-u]	(articolo m.)	321 (27)	830 (75)	59 (10)	69 (4)
/-u/	[-o]	/ˈe ˈɛ ˈo ˈɔ/	425 (12)	987 (114)	75 (6)	70 (2)
	[-u]	/ˈu ˈa ˈV + C palatale/	333 (30)	816 (36)	74 (7)	69 (6)

Tabella 4: Valori medi e deviazione standard di /u/ e /o/ toniche e finali atone (PSp)

In tutti e tre gli stili la vocale [-u] dell'articolo maschile mostra valori più vicini a quelli della [u] tonica che della [o] tonica e atona finale; parallelamente la [-o] dell'articolo neutro mostra valori tipici di [-o]. La situazione appare invece differenziata in base allo stile per la /-u/ delle altre parole lessicali. Nel parlato in isolamento, infatti, i valori formantici medi sono sempre sovrapponibili a quelli di [-u] dell'articolo maschile, quale che sia la qualità della vocale tonica; diversamente, entro frase e in parlato (semi)spontaneo, mostrano valori tipici di [-u], sovrapponibili a quelli di [-u] dell'articolo maschile, se in parole con vocale tonica non media, ma valori tipici di [-o] se in parole con vocale tonica media.

6.1 Parole funzionali (articolo, pronome clitico oggetto e dimostrativo): opposizione morfologica ab origine

La figura 1 evidenzia la netta distinzione tra le aree di [-u] e [-o] finali dell'articolo maschile e neutro, l'una attorno alla [u], l'altra attorno alla [o] tonica.

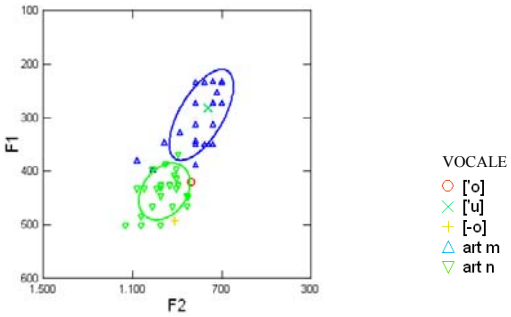


Figura 1: Dispersione di /-u/ e /-o/ nell'articolo (PI)

La figura 2 evidenzia una maggior area di dispersione e un avvicinamento delle aree di [-u] e [-o] finali dell'articolo maschile e neutro, che presentano un ristretto margine di sovrapposizione, rimanendo comunque sia attorno ai valori rispettivamente di [u] e [o] toniche.

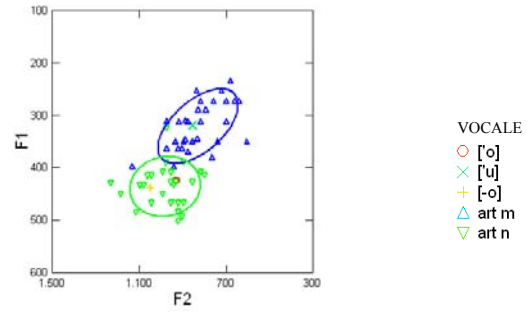


Figura 2: Dispersione di /-u/ e /-o/ nell'articolo (PF)

La figura 3 evidenzia la vicinanza del *corpus* relativo al parlato (semi)spontaneo al parlato entro frase piuttosto che a quello in isolamento. Le aree di esistenza di [-u] dell'articolo maschile e di [-o] dell'articolo neutro, nonostante il ristretto margine di sovrapposizione, si collocano attorno ai valori di [u] e [o] toniche.

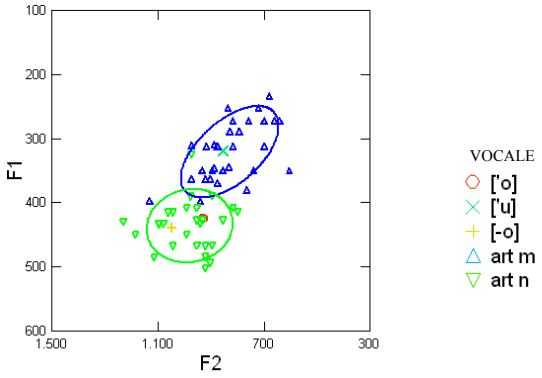


Figura 3: Dispersione di /-u/ e /-o/ nell'articolo (PSp)

6.1.1 Confronto tra stili

Anche ad una analisi di tipo uditivo la lista di frasi presenta cospicui fenomeni di coarticolazione, e la maggior vicinanza al parlato (semi)spontaneo che al parlato in isolamento.[6] Lo mostrano le frasi qui di séguito riportate:

(9) a. [ɔ ˈk̟o̟tto lu ˈp̟u̟jjo aˈrustu e lu k̟apˈpo bbolˈliːt̟u] {audio 2}
'ho cotto il pollo arrosto e il cappone bollito'

b. [ˈkwɛʃto ˈviːnu ˈɛ bˈboːnʊ] {audio 3}
'questo vino è buono'

Le informazioni relative alle parole funzionali sono sintetizzate in figura 4, ove sono riportati i valori medi delle [-u] e delle [-o] dell'articolo nelle tre tipologie di parlato, insieme con i valori medi di [u] tonica, [o] tonica, [-o] finale del parlato in isolamento, come termini di paragone. La rappresentazione grafica mostra chiaramente che, pur con la loro maggior dispersione e lieve sovrapposizione, i valori formantici medi di [-u] e [-o] negli articoli sono molto stabili e distinti tra loro, in tutte e tre le tipologie di parlato.

[6] Per l'italiano e le sue varietà v. ad esempio Savy & Cutugno (1997), Calamai (2001 [2005], 2003), Calamai & Sorianello (2003).

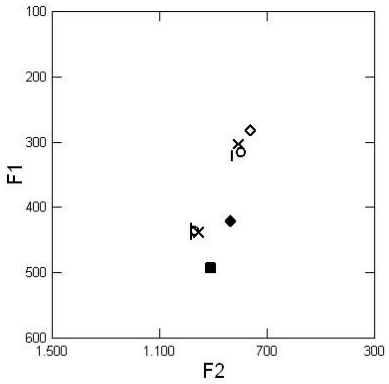


Figura 4: Valori medi PI (=X), PF (=O) e PSp (=|). Parole funzionali
◊ [ˈu] (PI) ♦ [ˈo] (PI) ■ [-o] (PI)

Ricapitolando, nel sistema dell'articolo, come nei pronomi tonici e atoni, e nei dimostrativi, la distinzione etimologica *-u* vs. *-o* risulta ancor oggi saldamente conservata in tutte e tre le tipologie di parlato.

6.1.2 Un confronto con i dialetti campani

La saldezza nel mantenimento dell'opposizione timbrica negli articoli e nei clitici è un dato in qualche modo sorprendente. Queste parole funzionali sono infatti monosillabiche e per di più atone. È quindi legittimo attendersi fenomeni di variabilità vocalica, correlati con la prominenza prosodica che ricevono in base al contesto segmentale e soprasegmentale in cui sono realizzate. È quanto avviene, ad esempio, nei dialetti campani, ove le ricerche di Maturi & Schmid (1999, 2001, 2002, 2003) hanno riscontrato l'esistenza di una correlazione tra i parametri soprasegmentali di durata e intensità e il timbro dei foni in esame.[7] Nei dialetti campani, sia per l'articolo determinativo maschile singolare sia per quello neutro, troviamo, in variazione apparentemente del tutto libera, le due varianti [o] e [u]. Dall'indagine dell'oscillazione tra [o] e [u] Maturi & Schmid (2002: 28) concludono che

> "le realizzazioni più brevi e più deboli presentano la massima variabilità di timbro, mentre quelle più lunghe e più intense presentano un timbro medio-alto; in altri termini si può ipotizzare che la forma lenta o profonda di questi morfi sia [o]".

In queste varietà, tuttavia, come mostrano gli esempi, la differenziazione di genere è garantita dalla presenza del raddoppiamento fonosintattico nel caso del neutro, e dalla sua

[7] Nel caso delle vocali posteriori, l'oscillazione tra il grado di apertura medio e quello alto caratterizza le seguenti parole funzionali monosillabiche: i) articoli determinativi maschile e neutro singolare; ii) clitici oggetto diretto maschile e neutro singolare; iii) articolo indeterminativo maschile; iv) negazione 'non'; v) preposizione 'con'; vi) bisillabo 'come'. Nel caso delle vocali anteriori, l'oscillazione caratterizza: i) articoli determinativi plurali maschile e femminile; ii) clitici oggetto diretto plurali maschile e femminile; iii) equivalente della preposizione 'di'.

assenza nel maschile (singolare): "p.es. nap. [o ˈvekə] 'lo vedo = vedo lui' vs. [o ˈvvekə] 'lo vedo = vedo ciò'" (Maturi & Schmid 2002: 26 n. 4].

Nel matelicese, invece, la differenziazione di genere maschile vs. neutro è garantita solo dalla distinzione di timbro *o* vs. *u*. La funzione morfologica, dunque, blocca la variazione libera tra i due timbri. Né gioca alcun ruolo la maggior consistenza fonica delle forme coinvolte, vale a dire il fatto che esse contengano anche una consonante. In gioco è il sistema.

6.2 Participio passato, aggettivo e nome: diffrazione di esiti dalla base -ŭ(m)

Andiamo ora a considerare le altre categorie lessicali, in cui ritroviamo la base etimologica -Ŭ(M), stavolta non opposta ad una base *-O, dunque senza connessione con il marcamento di genere. In nessuna categoria lessicale la disaggregazione secondo il genere ha prodotto risultati. Il livellamento analogico di marcamento del genere avvenuto a Macerata è dunque inesistente a Matelica.

Che la spiegazione della variabilità non possa esser esaurita in un richiamo all'influsso dell'italiano[8] mostra chiaramente la realizzazione della frase in (10), ove [-u] si conserva dopo vocale tonica non media, mentre tende ad abbassarsi ad [-o] dopo vocale tonica media:

(10) [lʊ maˈestrʊ a stuˈðiaːʊ̯u lo ˈɣrɛːko ˈsiː anˈtiːɻ̟u ɻ̟e mmoˈðɛrno] {audio 4}
'il maestro ha studiato il greco sia antico che moderno'

Descrivendo lo stesso fenomeno nella varietà di Vallepietra, Giancarlo Schirru (2009) parla di "diffusione della specificazione negativa [-alto]" a partire dalla schematizzazione illustrata in (11):

(11)

		caso maggioritario	caso minoritario
	[+alto]	/i u/ vocali di innesco	/u/ vocale bersaglio
[-basso]	[-alto]	/e o/ vocali bersaglio	/e ɛ o ɔ/ vocali di innesco
[+basso]		/a/ vocale neutra	/a/ vocale neutra

Interessante è il trattamento di /a/ come vocale neutra. Comparativamente, infatti, la vocale finale seguente /ˈa/ esibisce coinvolgimento variabile, diverso sia rispetto alle vocali finali dopo vocale tonica alta sia rispetto alle vocali finali dopo vocale tonica media (cfr. il tipo di variazione che indirizza, nella sua progressione, l'innalzamento campidanese descritto da Loporcaro, 2003). A Matelica la /ˈa/ si comporta come vocale neutra: non opera nessun condizionamento coarticolatorio rispetto alla situazione etimologica originaria.

Nella figura 5, relativa alle parole in isolamento, si osserva la sovrapposizione delle aree delle due classi di /-u/ finale, che occupano lo spazio attorno alla [u] tonica, ben distinto da quello di [o]. Gli assetti acustici delle parole isolate riflettono dunque la categorizzazione come /u/ nel giudizio dei parlanti.

[8] Che nel matelicese attuale agiscano anche effetti di 'italianizzazione' mostra in particolare NaC, l'informatore più giovane.

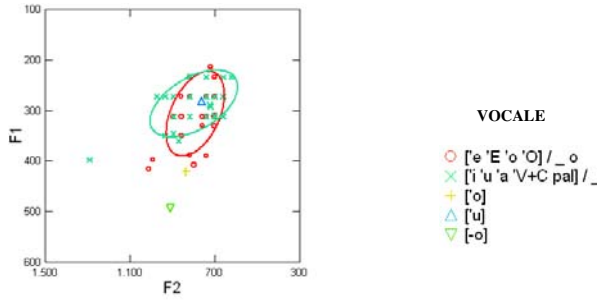


Figura 5: Dispersione di /-u/ atona finale nelle parole lessicali (PI)

La figura 6, relativa alle parole entro frase, evidenzia la maggior dispersione della /-u/ in parole con vocale tonica non media, o media seguita da (semi)vocale o (semi)consonante palatale, comunque sempre attorno all'area di [u] tonica. Dal diagramma spicca lo spostamento verso l'area di [o] dell'ellisse di /-u/ in parole con vocale tonica media.

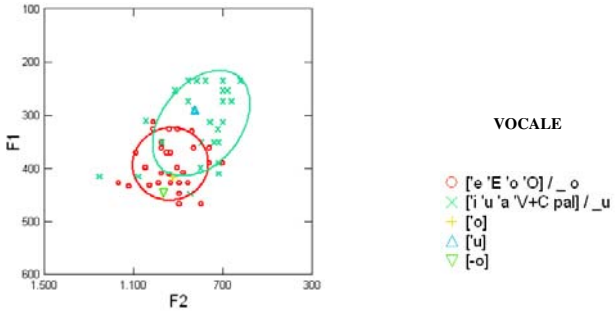


Figura 6: Dispersione di /-u/ atona finale nelle parole lessicali (PF)

La figura 7, relativa al parlato (semi)spontaneo, mostra un'ampia sovrapposizione tra le due aree di /-u/; si evidenzia inoltre una maggiore dispersione dell'area di /-u/ in parole con vocale tonica media. La variazione si orienta sia lungo l'asse verticale (apertura/chiusura) sia lungo quello orizzontale, con centralizzazione dell'elemento.

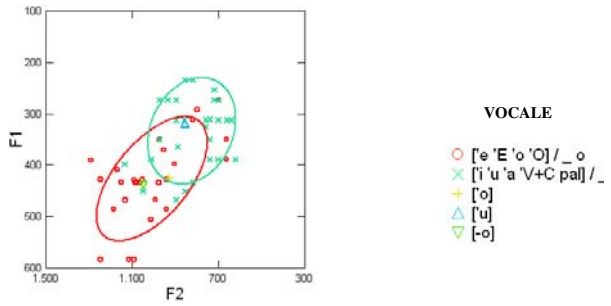


Figura 7: Dispersione di /-u/ atona finale nelle parole lessicali (PSp)

Di fronte a questi dati risulta confermata l'ipotesi che ci sia variazione di /-u/, e che questa variazione sia determinata dal contesto segmentale.

6.2.1 Confronto tra stili

Il confronto tra i valori medi relativi alle tre tipologie di parlato (fig. 8) evidenzia una bipartizione del materiale sonoro analizzato: la realizzazione delle parole in isolamento è più accurata e occupa spazi acustici più periferici, mentre i valori formantici medi provenienti dal vocalismo delle parole entro frase e del parlato (semi)spontaneo presentano minore perifericità e sono ampiamente sovrapponibili.[9]

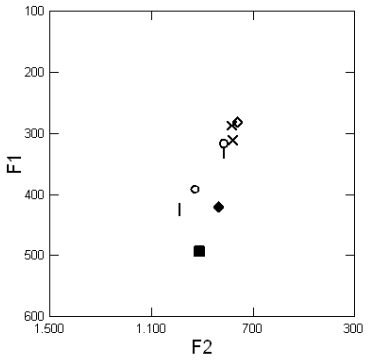


Figura 8: Valori medi PI (=X), PF (=O) e PSp (=|). Parole lessicali

◊ [ˈu] (PI) ♦ [ˈo] (PI) ■ [-o] (PI)

I valori formantici medi del vocalismo estratti dalle frasi si collocano all'interno dei valori medi del vocalismo estratti dal parlato in isolamento e all'esterno dei valori medi del vocalismo estratti dal parlato spontaneo.

[9] Questi dati confermano riflessioni ampiamente attese (cfr. almeno Calamai & Sorianello, 2003).

7. CONCLUSIONI

Risulta dunque dimostrato che l'abbassamento di /-u/ che oggi si presenta a Matelica è un fenomeno di variazione soggetto a condizioni coarticolatorie, senza ripercussioni sul sistema, perciò non accessibile alla coscienza metalinguistica del parlante, da ritenere dipendente dalla tipologia di parlato. L'indagine ha anche mostrato che la variazione non è generalizzata, ma subisce un condizionamento morfologico: essa è, infatti, ristretta alle sole parole lessicali, mentre non riguarda le parole funzionali, ove l'opposizione /-u/ ≠ /-o/ svolge la funzione morfologica di marcamento di genere maschile ≠ neutro.

La situazione di Matelica, raffrontata con il quadro di variazione dialettale della zona maceratese (e più ampiamente mediana), fornisce una conferma empirica del tipo di evoluzione che ha portato alla realizzazione di sistemi con armonia vocalica come quello di San Severino M. Il vocalismo atono finale del dialetto di San Severino è, infatti, il punto di arrivo di una vicenda che è partita da una configurazione etimologica di tipo reatino, con realizzazione di /-u/ univoca ([-u]), ed è stata determinata dalla coarticolazione a distanza da vocale a vocale.

RINGRAZIAMENTI

Ringrazio Paolo Bravi, Giovanna Marotta, Carlo Schirru, Stephan Schmid per le osservazioni espresse in occasione del convegno. Grazie inoltre ai tre anonimi giudici per i commenti. Esprimo la mia gratitudine agli amici matelicesi che si sono gentilmente prestati a rispondere alle mie domande: Giuseppe Crescentini, Nicola Crescentini, Domenico Crescentini. Grazie a Michela Traballoni, senza il cui aiuto queste persone sarebbero rimaste per me irreperibili. Grazie inoltre a Fabiola Branchesi e Giovanni Fiorani per la costruttiva discussione intorno ai problemi sollevati dall'analisi percettiva. Grazie a Marina Pucciarelli, Agostino Regnicoli e Mauro Salvatelli per l'assistenza tecnica.

8. BIBLIOGRAFIA

Baldini, T. (2006), Lu bagnu al mare, in *"... Lì comincia 'na vallata che pare un budéllu ...". Testimonianze dialettali delle Alte Valli del Potenza e dell'Esino* (M. Pucciarelli & A. Regnicoli, editors), San Severino Marche: Biemmegraf, 88-89.

Balducci, S. (2000), *Marche*, Pisa: Pacini.

Biondi, A. (2003), *Materiali per il* Vocabolario del dialetto di San Severino Marche, Ms.

Boldrini, V. (2006), Aurora, in *"... Lì comincia 'na vallata che pare un budéllu ...". Testimonianze dialettali delle Alte Valli del Potenza e dell'Esino* (M. Pucciarelli & A. Regnicoli, editors), San Severino Marche: Biemmegraf, 67.

Bricchi, A. (1984), *Matelica. I suoi abitanti. Il suo dialetto*, Matelica: Associazione Pro Matelica.

Calamai, S. (2001 [2005]), Stili a confronto nel parlato toscano (Pisa e Firenze), *ID*, 65, 95-125.

Calamai, S. (2003), Vocali d'Italia. Una prima rassegna, in *Scritti in onore di Franco Ferrero* (P. Cosi, editor), Padova: Unipress, 49-58.

Calamai, S. & Sorianello, P. (2003), Aspetti stilistici del vocalismo romano, *Quaderni del Laboratorio di Linguistica della Scuola Normale Superiore di Pisa*, 4 (n.s.), 27-41.

Camilli, A. (1929), Il dialetto di Servigliano, *AR*, 13, 220-271.

Campanelli, B. (1896), *Fonetica del dialetto reatino ora e per la prima volta studiata sulla viva voce del popolo*, Torino: Loescher.

Contini, G. (1961-1962), Clemente Merlo e la dialettologia italiana, *AATSL* 26 (n.s. 12), 325-341 [poi in ID., *Altri esercizî (1942-1971)*, Torino: Einaudi, 1972, 355-367].

Gebhardt, T. (1997), *La regressione dialettale a Matelica*, MThesis, Univ. of Heidelberg, Germany.

Goldsmith, J.A. (1995), *The Handbook of Phonological Theory*, Cambridge, MA: Blackwell.

Harris, J. (1994), Monovalency and opacity: Chichewa height harmony, *UCL Working Papers in Linguistics*, 6, 509-545.

Leopardi, A. (1887), *Sub tegmine fagi. Sotto un tegame di fagiuoli*, Città di Castello: S. Lapi.

Lindstrom, A. (1907), Il vernacolo di Subiaco, *Studj romanzi*, 5, 237-300.

Loporcaro, M. (2003), Coarticolazione e regolarità del mutamento: l'innalzamento delle vocali medie finali in sardo campidanese, in *La coarticolazione*. *Atti delle XIII Giornate di Studio del Gruppo di fonetica sperimentale, Pisa, 28-30 novembre 2002* (G. Marotta & N. Nocchi, editors), Pisa: ETS, 23-44.

Lorenzotti, A. (2000), *Dialetto, cultura e saperi tradizionali a Matelica*, MThesis, Sapienza Univ. of Rome, Italy.

Maiden, M. (1988), Armonia regressiva di vocali atone nell'Italia meridionale, *ID*, 51, 111-139.

Maiden, M. (1989), Sulla morfologizzazione della metafonesi nei dialetti italiani meridionali, *Zeitschrift für romanische Philologie*, 105, 178-192.

Maiden, M. (1991), *Interactive morphonology. Metaphony in Italy*, London: Routledge.

Maiden, M. (1995), Evidence from the Italian dialects for the internal structure of prosodic domains, in *Linguistic Theory and the Romance Languages* (J.C. Smith & M. Maiden, editors), Amsterdam: Benjamins, 115-131.

Maturi, P. & Schmid, S. (1999), Phonetically conditioned allomorphy of functional words in a dialect of Southern Italy, in *Proceedings of the 14th International Congress of Phonetic Sciences*, Berkeley: University of California, 1393-1396.

Maturi, P. & Schmid, S. (2001), Allomorfia e morfo-fonetica: riflessioni induttive su dati dialettali campani, in *Dati empirici e teorie linguistiche* (F. Albano Leoni, R. Sornicola, E. Stenta Krosbakken & C. Stromboli, editors), Atti del XXXIII Congresso Internazionale di Studi della Società di Linguistica Italiana, Napoli, 28-30 ottobre 1999, Roma: Bulzoni, 251-265.

Maturi, P. & Schmid, S. (2002), Dialettologia e fonetica acustica. Una ricerca in Campania, in *La fonetica acustica come strumento di analisi della variazione linguistica in Italia. Atti delle XII Giornate di Studio del Gruppo di Fonetica Sperimentale,* Macerata, 13-15 dicembre 2001 (A. Regnicoli, editor), Roma: il Calamo, 23-28.

Maturi, P. & Schmid, S. (2003), Sulla diffusione areale di un fenomeno di variazione morfo-fonetica nei dialetti campani, in *Actas del XXIII Congreso Internacional de Lingüística y Filología Románica*, Salamanca, 24-30 septiembre 2001 (F. Sánchez Miret, editor), Tübingen: Niemeyer, 221-233.

Mengel, E. (1936), *Umlaut und Diphthongierung in den Dialekten des Picenums*, PhDThesis, Univ. of Köln, Germany.

Merlo, Cl. (1906-1907), Dei continuatori del lat. ILLE in alcuni dialetti dell'Italia centro-meridionale, *ZrPh*, 30, 11-25, 438-454; 31, 157-163.

Merlo, Cl. (1920), Fonologia del dialetto di Sora (Caserta), *Annali delle Università Toscane* 38, 117-283 [poi vol. a sé, Pisa: Mariotti, 1920].

Merlo, Cl. (1922), *Fonologia del dialetto della Cervara in provincia di Roma*, Roma: Società Filologica Romana.

Merlo, Cl. (1930), *La donna di Guascogna e il re di Cipro. Novella (Decam I, 9) tradotta nei parlari del Lazio. I. Valle dell'Aniene,* Roma: Società Filologica Romana.

Meyer-Lübke, W. (1890), *Italienische Grammatik*, Leipzig: Reisland.

Paciaroni, T. (in stampa), Verso l'armonia vocalica. Diffrazione degli esiti di -/u/ nel dialetto di Matelica, *in Lingue, ETHNOS E Popolazioni: evidenze linguistiche, biologiche e CULTURALI*, Atti del XXXII Convegno annuale della Società Italiana di Glottologia (S.I.G.), Verona, 25-27 Ottobre 2007 (P. Cotticelli, editor).

Paciaroni, T. & Loporcaro, M. (in stampa), Funzioni morfologiche della distinzione fra *-u* e *-o* nei dialetti del Maceratese, in *Actes du XXV Congrès International de Linguistique et de Philologie Romanes*, Innsbruck, 3-8 septembre 2007 (M. Iliescu, H. Siller & P. Danler, editors), Tübingen: Niemeyer.

Parrino, F. (1967), Per una carta dei dialetti delle Marche, *BCDI* 2, 7-37.

Pucciarelli, M. (2006), Fenomeni di armonia vocalica nel dialetto maceratese, *Rivista italiana di linguistica e di dialettologia*, 8, 87-114.

Rohlfs, G. (1949), *Historische Grammatik der Italienischen Sprache und ihrer Mundarten. II. Formenlehre und Syntax*, Bern: Francke (trad. it. *Grammatica storica della lingua italiana e dei suoi dialetti. Morfologia*, Torino: Einaudi, 1968).

Salvioni, C. (1900), Il 'Pianto delle Marie' in antico volgare marchigiano, *Reale Accademia dei Lincei*, 8, 577-605.

Savy, R. & Cutugno, F. (1997), Ipoarticolazione, riduzione vocalica, centralizzazione: come interagiscono nella variazione diafasica?, in *Fonetica e fonologia degli stili dell'italiano parlato,* Atti delle VII Giornate di Studio del Gruppo di Fonetica Sperimentale, Napoli, 14-15 Novembre 1996 (F. Cutugno, editor), 177-194.

Schirru, G. (2009), Osservazioni sull'armonia vocalica nei dialetti della Valle dell'Aniene, Comunicazione alle Giornate di Studio *Vicende storiche della lingua di Roma*, Zurigo, Università di Zurigo, 17-19 settembre 2009.

Traballoni, M. (2002-2003), *Glossario del dialetto di Matelica*, MThesis, Univ. of Macerata, Italy.

Vignuzzi, U. (1988), Italienisch: Areallinguistik, VII. Marche, Umbrien, Lazio, in *Lexikon der Romanistischen Linguistik* (G. Holtus, M. Metzeltin & Ch. Schmitt, editors), Vol. IV, Tübingen: Niemeyer, 606-642.

DURATA E STRUTTURE FORMANTICHE NEL PARLATO TOSCANO: INDAGINI PRELIMINARI SU UN CAMPIONE DI DIALOGHI SEMISPONTANEI*

Nadia Nocchi [a], Silvia Calamai [b]
[a] Università di Zurigo, [b] Università degli Studi di Siena (sede di Arezzo)
nocchi_nadia@yahoo.com, calamai@unisi.it

1. SOMMARIO

A partire da un *corpus* articolato e particolarmente complesso anche in considerazione della sua diversificazione geografica 'fine' (da dialettologi), il contributo analizza il vocalismo del parlato di area toscana osservando un parametro prosodico (la durata) e alcuni parametri segmentali (le prime due formanti). L'architettura del disegno sperimentale mira a fornire un contributo in merito sia a interrogativi di fonetica generale, sia a interrogativi di carattere più squisitamente geolinguistico, essendo tre le città toscane di cui si presentano i dati acustici. Il materiale sonoro analizzato è di tipo semispontaneo (dialoghi *map task*): sebbene sia stato raccolto con un medesimo protocollo sperimentale nelle tre località, pone inevitabilmente questioni cruciali sia di rappresentatività statistica, sia di difficoltà oggettive nel rilevamento dei valori formantici. Anche per questa ragione, ad osservazioni quantitative affianchiamo, nella seconda parte, considerazioni squisitamente qualitative.

In primo luogo il *corpus* è indagato, da un punto di vista quantitativo, nella sua interezza: viene esplorata l'influenza di alcuni fattori linguistici – 'accento' (accento di parola *vs.* accento di frase), 'posizione' (iniziale, interna, finale di enunciato) e, più parzialmente, 'tipo di parola' – nella distribuzione dei valori formantici e temporali per il vocalismo tonico. I fattori 'posizione' e 'tipo di parola' vengono poi osservati nel vocalismo atono. Infine, il vocalismo tonico viene confrontato con quello atono. In secondo luogo sono osservati i valori formantici e temporali suddivisi per località, al fine di indagare la possibile influenza del fattore 'spazio' sia nel vocalismo tonico che in quello atono. Per il sistema tonico scorporato per località si fornisce anche un quadro relativo all'influenza del fattore 'accento'. Un confronto tra vocalismo tonico e vocalismo atono nei tre punti d'indagine chiude l'analisi quantitativa. In conclusione, viene proposto un tentativo di ispezione fine dell'evoluzione temporale dei movimenti formantici all'interno di un numero ristretto di dati – sempre suddivisi per località – per capire se gli aspetti dinamici dei segmenti vocalici permettono, anch'essi, di rendere conto di alcune differenze geolinguistiche.

Dal momento che il quadro del vocalismo toscano – anche nei suoi aspetti più tipizzanti – ci è noto, dalla letteratura, negli stili più controllati (parlato letto di varie tipologie), il presente contributo permette anche di valutare in quale misura viene preservata la tipicità del vocalismo pisano e soprattutto livornese (abbassamento delle vocali medio-basse e

*Per finalità accademiche italiane: NN è responsabile della stesura dei §§ 3 e 5; SC dei §§ 2, 4.1, 4.2, 4.4, 6. Il § 4.3 è stato scritto congiuntamente dalle due autrici. Le misurazioni del campione livornese sono state compiute da NN, quelle del campione fiorentino sono state fatte da entrambe le autrici, quelle del campione pisano sono ad opera di SC. SC ha curato la parte statistica del lavoro; NN ha predisposto tutte le rappresentazioni grafiche.

posteriorizzazione di [a]) in stili di eloquio più liberi (parlato semispontaneo). I dati riferiti al vocalismo atono – per la prima volta presentati su un campione di parlato semispontaneo di base toscana – intendono offrire un contributo in merito al controverso tema che va sotto il nome di 'riduzione vocalica'.

2. DURATA E SOTTOSPECIFICAZIONE ACUSTICA: DALLA FONETICA ALLA SOCIOFONETICA

Si è soliti distinguere, in letteratura, due tipologie differenti di riduzione vocalica: una di tipo fonetico, dal carattere gradiente, legata a fattori stilistici quali la velocità d'eloquio e la maggiore o minore ipoarticolazione, connessa anche a fattori più strettamente linguistici quali ad esempio la tipologia della parola (rispetto alle parole lessicali, quelle funzionali sono solitamente soggette a riduzione maggiore) e la frequenza d'uso (parole più frequenti sono solitamente più ridotte di parole meno frequenti); l'altra, di carattere fonologico e strutturale, legata al passaggio dal sistema tonico al sistema atono, in cui avviene una neutralizzazione di alcuni contrasti fonologici, spesso mediante lo *schwa*. Si tratta in questo caso di una sostituzione categoriale che non dipende né dalla velocità d'eloquio né dallo stile. Per i due tipi di riduzione sono state proposte diverse definizioni: *phonetic vowel reduction* vs. *phonological vowel reduction* (Fourakis, 1991); *acoustic vowel reduction* vs. *lexical vowel reduction* (van Bergem, 1993; 1995); 'riduzione timbrica' o 'riduzione non strutturale' vs. 'centralizzazione strutturale' o 'riduzione strutturale' (Savy & Cutugno, 1997; Lo Prejato *et al.*, 2004; Clemente, 2005).

Il tema della riduzione vocalica intesa come sottospecificazione acustica (*vowel reduction as target undershoot*) ha come primo, imprescindibile, punto di riferimento il modello di Lindblom (1963) secondo cui, in presenza di riduzioni temporali, gli organi articolatori non sono in grado di raggiungere il bersaglio previsto. L'analisi acustica di otto vocali svedesi, toniche e atone, a parlato lento e veloce, ha dimostrato che parlato veloce e assenza di accento producono un accorciamento temporale, e che questo risulta essere fortemente correlato alla sottospecificazione formantica, definita come incapacità di raggiungere il 'bersaglio' (*target*) della frequenza formantica per una data vocale. In questo quadro, la centralizzazione può essere uno tra gli esiti acustici possibili, ma non l'unico: la riduzione si presenta come un fenomeno assimilatorio, poiché le formanti vengono alterate in base all'ambiente segmentale circostante. Una conseguenza della riduzione è inoltre un generale restringimento dello spazio vocalico, ovvero una più ridotta distanza tra vocali (v. Fourakis, 1991 e van Bergem, 1993, rispettivamente per l'inglese americano e l'olandese). In ogni caso, le conseguenze fonetiche della riduzione vocalica sono ancora oggetto di dibattito scientifico: per taluni si tratta di una complessiva centralizzazione verso la posizione dello *schwa*, per altri sono forme di assimilazione contestuale.[1]

Gli effetti della riduzione vocalica sono stati osservati negli anni in molteplici condizioni e sotto differenti punti di vista: presenza *vs.* assenza di accento, parlato lento *vs.* parlato veloce, parlato letto *vs.* parlato spontaneo, parole lessicali *vs.* parole funzionali, forme elicitate in modalità cosiddetta iperarticolata (*clear*) *vs.* forme di citazione.[2] L'ipotesi

[1] Nella sintesi di Mooshammer & Geng, (2008: 118-119) le due diverse posizioni sono definite, rispettivamente, *paradigmatic reduction* e *syntagmatic reduction*.

[2] Si veda la rassegna in Rosner & Pickering, (1994), le pagine introduttive di Padgett & Tabain (2005) e di Mooshammer & Geng (2008).

di lavoro più semplice (per certi versi semplicistica), non sempre confermata dalle analisi sperimentali, prevede che il massimo della riduzione si verifichi in assenza di accento, nel parlato veloce, in quello spontaneo, in parole funzionali; mentre il minimo della riduzione sia atteso in presenza di accento, nel parlato lento, in quello letto, nelle parole lessicali e nelle forme iperarticolate, che in quanto tali, tendono a sfruttare al massimo lo spazio vocalico disponibile. In realtà il modello di Lindblom, (1963) è stato, nel tempo, sottoposto a diverse critiche e ha subìto da parte dello stesso autore alcune revisioni che limitano la capacità della durata nel predire gli andamenti formantici (Moon & Lindblom, 1994; van Bergem, 1993; van Bergem, 1995; van Son & Pols, 1990; 1992). Il grado di dipendenza dalla durata da parte dei valori formantici sembra variare sia in base a fattori stilistici in parte attivamente controllati dal parlante, sia in base a fattori squisitamente idiosincratici, dal momento che i singoli parlanti risultano essere in grado di variare il grado di sottospecificazione acustica a prescindere da fattori stilistici: "Some speakers show very strong effects, whereas others show a smaller trend" (Moon & Lindblom, 1994: 47). Per quanto concerne la velocità d'eloquio, le analisi di van Son & Pols, (1990) e di van Son, (1993) hanno evidenziato la mancanza di un evidente *undershoot* in relazione all'aumento della velocità: è possibile che a velocità più elevate entrino in gioco vere e proprie riorganizzazioni delle strategie articolatorie, in grado di raggiungere ugualmente i bersagli acustici. Un differente stile di eloquio (veloce *vs.* lento) può cambiare la durata delle vocali senza alterare i valori formantici, oppure può cambiare i valori formantici in una maniera inaspettata (van Son & Pols, 1990: 1692). Anche usando vocali in contesti identici, una semplice correlazione tra formanti e durata non può essere estesa a stili di eloquio differenti; pertanto il potere esplicativo della durata nel predire la posizione del bersaglio vocalico deve essere giudicato marginale, se confrontato con altri fattori di carattere contestuale. Infine, il grado di *undershoot* pare essere legato anche al tipo di vocale in questione (la variazione è maggiore nelle formanti di certe vocali rispetto ad altre): è stato sperimentalmente dimostrato, ad esempio, come le frequenze formantiche varino di più nelle vocali rilassate piuttosto che in quelle tese, più stabili (Picheny, Durlach & Braida, 1986: 441-443; Moon & Lindblom, 1994: 47).

Per quanto riguarda la variabile accento, il vocalismo tonico presenta generalmente vocali più periferiche rispetto a quello atono (cfr., per l'inglese americano, Fourakis, 1991) e lo spazio vocalico appare nel complesso di minore estensione (per l'italiano in stili non controllati v., tra gli altri, Savy & Cutugno, 1997). Tuttavia, se la riduzione vocalica è intesa come un fenomeno di centralizzazione, lo spazio vocalico viene ridotto verso il centro, se invece è considerata come assimilazione contestuale esso in qualche modo si modella sul luogo di articolazione delle consonanti adiacenti. Ad ogni buon conto, in questa sede i diversi modelli interpretativi non sono osservati nei loro aspetti predittivi – a nostro avviso ancora controversi – ma sono indagate le possibili implicazioni geolinguistiche dei diversi andamenti temporali e timbrici.

Le indagini sul *formant undershoot* sono state generalmente svolte in contesti controllati a causa dell'oggettiva difficoltà di modellizzare tutte le variabili in gioco; negli stili più liberi gli effetti dovuti all'*undershoot* potrebbero essere erroneamente etichettati come effetti sociolinguistici (stilistici, diatopici) mentre sono solo effetti fonetici; viceversa, gli aspetti più squisitamente geolinguistici possono venire mascherati dagli effetti meramente fonetici. Il punto è cruciale per un'analisi fonetica che abbia, tra i suoi obiettivi, anche una prospettiva di tipo più sociolinguistico. L'area toscana è a nostro avviso un laboratorio

privilegiato per una riflessione anche metodologica su questioni, appunto, di sociofonetica, dal momento che il quadro acustico, per alcune località della regione, è relativamente noto, almeno sul materiale controllato (lettura di liste di parole, lettura di frasi).[3] La diversità rilevata – laddove presente – potrebbe con una certa sicurezza essere imputata a fattori diatopici.

Il tipo di materiale utilizzato per la nostra ricerca, qui presentata in una forma preliminare, differenzia la nostra analisi dai capisaldi della letteratura internazionale, le cui analisi in genere sono basate su *corpora* estremamente controllati ed elicitati in laboratorio, con parlanti spesso fonetisti essi stessi. D'altra parte, le indagini condotte in area italiana che prendono in esame anche campioni diversificati in diatopia non privilegiano una prospettiva sociofonetica (v. ad es. Savy *et al.*, 2005) che invece costituisce il filo conduttore del nostro percorso di ricerca.[4]

L'analisi verterà su parlato relativamente spontaneo, dunque, per certi versi, già 'ridotto' in partenza, anche se caratterizzato da diversi gradi e diversi livelli di riduzione (per i dettagli v. § 2): una strutturale (tonico *vs.* atono), una prosodica (per le sole toniche, tonica di enunciato *vs.* tonica lessicale), una posizionale (posizione della vocale sotto esame nella parola e, nel caso delle vocali toniche, nella catena parlata), una semantica (parole piene *vs.* parole vuote)[5]. Questa architettura si ripete per tre differenti località (con alcune limitazioni che discuteremo in § 2): il *corpus* di parlato sotto esame appare dunque sociolinguisticamente omogeneo ma geolinguisticamente differenziato[6]. Inoltre, la vicinanza, teorica e metodologica, della dialettologia alla sociolinguistica – soprattutto in un territorio come quello italiano (Berruto, 1995; Calamai, 2007) – ci permette di adottare una prospettiva sociofonetica nell'analisi e nella discussione dei risultati.

[3] Per il vocalismo tonico pisano e livornese v. Calamai, (2004a), per un confronto con il vocalismo di Firenze v. Calamai, (2001; 2004b).

[4] Un importante e recente contributo in questo senso è Abete & Simpson (in questo volume). Precisiamo che utilizziamo il prefissoide *socio-* nel senso più ampio del termine (diatopico, diastratico, diafasico, diamesico) e che esula dagli obiettivi del presente lavoro un'analisi su soggetti diversificati secondo variabili diastratiche (età, livello socio-culturale).

[5] Con l'etichetta di 'parole piene' intendiamo le parole lessicali (soprattutto sostantivi e aggettivi, buona parte dei verbi); con l'etichetta di 'parole vuote' intendiamo le parole funzionali (ovvero congiunzioni, preposizioni, articoli, parte degli avverbi, i verbi ausiliari…).

[6] Sullo sfondo resta il problema del confronto interlocutore: la questione non è di poco conto visto che nonostante l'omogeneità di raccolta dei dati e la tipologia di parlato ottenuto (ricorrenza di medesime entrate lessicali quali *andare*, *girare*, *sinistra*, *destra*…) abbiamo a che fare con parlanti *diversi*, ciascuno con le proprie caratteristiche fisiche e le proprie idiosincrasie. Una possibile soluzione – sensibilmente differente dalle procedure 'tradizionale' di normalizzazione – è in Mooshammer & Geng (2008) proprio in riferimento alla riduzione vocalica.

3. MATERIALI E METODI

3.1 I parlanti

Il *corpus* preso in esame è stato ottenuto tramite la tecnica del *map task* (cfr. Brown *et al.*, 1984): ai due partecipanti sono state consegnate due mappe diverse, in cui soltanto una presenta un tracciato. Il compito dei due soggetti è stato quello di trasferire – quanto più accuratamente possibile – il percorso da una mappa all'altra tramite l'interazione verbale. Le due mappe non sono completamente identiche, poiché sia il numero, sia la natura e la posizione degli oggetti differiscono per qualche aspetto, in modo tale da rendere possibili eventuali fraintendimenti o momenti di difficoltà nello scambio comunicativo, così come può avvenire in situazioni di interazione reale. Il materiale acustico raccolto con questa tecnica costituisce un buon compromesso, dunque, tra due estremi, rappresentati dalle registrazioni nascoste e dal parlato letto di laboratorio. Inoltre, il metodo del *map task* permette sia di controllare il tipo di situazione comunicativa, sia di verificare quanto lo scambio comunicativo abbia avuto successo, valutandolo in termini di accuratezza nella trasposizione del percorso da una mappa all'altra.

Sono quattro le città toscane scelte per questa ricerca, spostandosi da ovest verso est: Livorno, Pisa, Firenze e Arezzo; di quest'ultima le misurazioni sono ancora in corso e non verranno presentate in queste sede. Se il metodo di rilevamento dei dati è lo stesso per tutte le località, diversi sono i progetti all'interno dei quali sono stati raccolti i dati. Per Pisa e Firenze si sono utilizzate le mappe create per il progetto A.P.I. (*Archivio di Parlato Italiano*).[7] Per Livorno e Arezzo i dati sono stati raccolti *ex novo* dalle due Autrici all'interno di progetti di ricerca personali. Il numero totale dei soggetti analizzati è pari a otto:[8] si tratta per la maggior parte di studenti universitari maschi di un'età compresa, al momento della registrazione, tra 20 e 30 anni. La raccolta dei dati è avvenuta in larga parte all'interno della camera anecoica del Laboratorio di Linguistica della Scuola Normale Superiore.[9]

---

[7] API è un progetto cofinanziato MIUR (cofin. 99), iniziato nel novembre 1999 e concluso nel novembre 2001. Ha avuto come finalità l'analisi segmentale, prosodica e testuale di una vasta porzione del materiale raccolto per il progetto A.V.I.P. (*Archivio delle Varietà di Italiano Parlato*). Il *corpus* utilizzato per il progetto A.P.I. comprende produzioni linguistiche semispontanee raccolte nelle aree di Pisa, Firenze, Napoli, Bari e Roma (cfr. http://www.parlaritaliano.it/parlare/dati_e_strumenti/41/API.php).

8 Per la precisione: due sono i soggetti pisani (*corpus* AVIP- API), due quelli fiorentini (uno del *corpus* AVIP- API e uno registrato dal primo autore) e quattro i parlanti livornesi (*corpus* raccolto dal primo autore).

[9] I soggetti aretini sono stati registrati all'interno del Laboratorio di Linguistica Sperimentale presso la Facoltà di Lettere di Arezzo. Per quanto riguarda la strumentazione tecnica, è stato utilizzato – per il campione livornese – un registratore Sony DAT (*Digital Audio Tape*) modello DTC-ZE 700, su cassette DAT TDK da 60 e da 90 minuti; per il campione aretino, un registratore digitale M-Audio Microtrack 24/96 e microfono professionale Shure SM58. Per motivi essenzialmente di tipo logistico, due mappe (una fiorentina e una livornese) sono state registrate all'interno di una abitazione, in un ambiente quindi non insonorizzato, ma comunque poco rumoroso, sempre con microfono professionale SONY ECM-719 e mediante registratore portatile Sony DAT modello TCD-D8 su cassette DAT TDK da 60 e 90 minuti.

3.2 Criteri di misurazione e architettura del corpus

L'analisi strumentale è stata effettuata mediante i *softwares* Praat (versione 4636) (cfr. Boersma & Weenink, 2005) e Multi-Speech usando la tecnica FFT[10] che, a differenza dell'inviluppo LPC, permette di rilevare anche valori formantici prossimi l'uno all'altro. Questi i parametri misurati: F1, F2, f0, e durata delle vocali toniche e atone dei quattro sistemi vocalici. La misurazione delle formanti è avvenuta in tre punti (poi mediati)[11] della parte centrale del segmento vocalico, Il numero complessivo di ricorrenze analizzato è pari a 864 per il vocalismo tonico e a 666 per le vocali atone. Nella tabella 1 si riportano i dati analizzati suddivisi per località:

località	***vocalismo tonico***	***vocalismo atono***
Firenze	311	238
Pisa	359	237
Livorno	194	191
totale	864	666

Tabella 1: Numero di ricorrenze misurate per ogni località

All'interno del *corpus*, le vocali toniche sono state indicizzate sulla base di tre fattori: 1) accento (tonica di enunciato *vs.* tonica lessicale); 2) posizione (iniziale, interna, finale di turno o prima di pausa); 3) tipo di parola (parola funzionale *vs.* parola lessicale). Per le atone sono stati considerati soltanto il fattore posizionale (interna di parola e finale lessicale ma interna di enunciato)[12] e il fattore semantico (parola funzionale *vs.* parola lessicale). L'etichettatura 'prosodica' è stata condotta in maniera impressionistica sulla base di valutazioni percettive, seguite dal controllo di due parametri acustici quali l'andamento della frequenza fondamentale e l'intensità (cfr. Marotta, Calamai, & Sardelli, 2004). Un'analisi più raffinata è in preparazione, soprattutto per quanto concerne il *corpus* livornese.

Il carattere 'spontaneo' del materiale raccolto ha influito sull'analisi acustica e sull'interpretazione dei risultati. In primo luogo, l'impossibilità di bilanciare, se non controllare, il contesto consonantico adiacente (nonostante il lessico usato sia piuttosto limitato, dato il tipo di compito richiesto ai locutori e la natura stessa del *map task*) rende piuttosto elevati gli intervalli di variazione di entrambe le formanti per ciascun fono; in secondo luogo, l'impossibilità di bilanciare numericamente – almeno in questa prima fase della ricerca – tutti i sottogruppi oggetto dell'indagine (ovvero, l'impossibilità di raggiungere una numerosità

[10] Queste le impostazioni per l'inviluppo spettrale FFT: w*indow length* 0.005 sec; *window shape:* hamming; *pre-emphasis- level:* 6 db.

[11] È stata da più parti sottolineata la difficoltà di individuare la cosiddetta porzione stabile (*steady state*) delle formanti nei foni vocalici, soprattutto nel parlato meno controllato (v. Rosner & Pickering, 1994: 279): la scelta del punto su cui compiere la misurazione non è del resto soltanto una questione operativa. Si è deciso per questa parte del lavoro di rilevare i valori nel punto centrale della vocale, dato dalla media dei tre valori presi a distanze equivalenti (al 45%, al 50% e al 55% della durata del segmento), in modo tale da limitare, per quanto possibile, gli effetti coarticolatori con i suoni consonantici adiacenti. Dalle misurazioni sono stati esclusi i dittonghi e gli iati.

[12] Le vocali atone finali prima di pausa sono state sì indicizzate ma escluse dalla presente analisi.

adeguata per tutte le vocali nelle tre diverse posizioni, o l'impossibilità di raggiungere un numero comparabile di vocali sia nella parole piene che nelle parole vuote) permette di sottoporre i dati a un'analisi statistica soltanto parziale. Questo sbilanciamento numerico si è rivelato particolarmente pesante per [u] che – vista anche la sua limitazione distribuzionale nella fonologia dell'italiano – è stata analizzata soltanto da alcuni punti di vista.

4. ANALISI QUANTITATIVA

La presentazione dei risultati segue un ordine che dal più generale (le vocali 'toscane' considerate nel loro complesso) arriva al particolare (le vocali di Firenze, Pisa, Livorno): in primo luogo, dunque, il *corpus* viene osservato in quanto *corpus* del parlato toscano *tout court* e gli effetti dei fattori linguistici presi in esame saranno indagati in tutte le vocali toscane (appunto in quanto vocali 'toscane');[13] in secondo luogo, l'ispezione prenderà in esame il fattore 'luogo' per rintracciare andamenti comuni ed eventuali specificità.

4.1 Gli effetti di 'accento', 'posizione', 'tipo di parola' sulle vocali di Toscana

L'elevata numerosità del campione ci permette di valutare statisticamente (con poche eccezioni) l'effetto dei tre fattori – accento lessicale *vs*. accento frasale (per le sole vocali toniche), posizione della vocale, tipo di parola – sulle due formanti e sulla durata. Nella presentazione dei risultati, i dati relativi alle vocali toniche sono seguiti da quelli relativi alle vocali atone (eccezion fatta, naturalmente, per il fattore accento).

Per quanto riguarda l'opposizione accento di parola *vs.* accento di frase, nella tabella 2 vengono riportati i dati complessivi e le significatività statistiche.

[13] Del resto, una omogeneità nel retroterra geografico relativamente ai locutori utilizzati nelle analisi fonetiche sperimentali non è così diffusa come un dialettologo potrebbe auspicare: non è infrequente trovare assembramenti di parlanti di provenienze anche molto diverse. Si veda ad esempio Mooshammer & Geng (2008: 122): "All speakers spoke a standard variety of German with slight dialectal variations: three speakers originally come from the south of Germany, one speaker from Saxony, two speakers from Northeast Germany and one speaker from Berlin", fino alla sorprendente mescidanza di Padgett & Tabain (2005: 20): "nine speakers of Russian were recorded […] speakers were aged between 19 and 64, and had spent between 1 and 44 years in Australia […] about half the speakers are not from Russia: 3 are from China, 1 is from Ukraine (Kiev), and another had spent time in Ukraine, Uzbekistan and Moscow". Ci sentiamo pertanto autorizzate a osservare le nostre vocali anche complessivamente come vocali 'toscane', eventualmente da confrontare, in futuro, con Ferrero (1972), con i dati fiorentini del *corpus* DIVA di Albano Leoni, Cutugno & Savy (1995) e con le ormai numerose indagini sui vocalismi 'regionali' della penisola italiana (si veda, tra gli altri, Schirru 1994; 2003).

v	*F1*			*F2*			*durata*		
	accento lessicale	accento frasale	valori di t	accento lessicale	accento frasale	valori di t	accento lessicale	accento lessicale	valori di t
a	650 (102)	716 (105)	-4.491 (213) ***	1353 (133)	1306 (113)	2.626 (213) **	94 (34)	134 (39)	-7.676 (213)***
ɛ	538 (84)	589 (83)	-3.939 (169) ***	1693 (142)	1762 (140)	-3.167 (169) **	84 (35)	137 (42)	-8.995 (169)***
e	422 (55)	390 (55)	2.603 (77)*	1941 (154)	1973 (153)	n.s.	69 (37)	85 (38)	n.s.
i	352 (62)	348 (72)	n.s.	2102 (172)	2186 (241)	-2.515 (151)*	71 (31)	103 (27)	-6.963 (151)***
ɔ	562 (70)	592 (91)	n.s.	1131 (156)	1080 (99)	n.s.	98 (60)	125 (46)	-2.367 (99)*
o	468 (72)	429 (50)	2.368 (74)*	1146 (228)	1003 (192)	2.632 (74)**	77 (35)	115 (32)	-4.471 (74)***
u	399 (72)	379 (69)	n.s.	957 (201)	925 (181)	n.s.	77 (34)	93 (38)	-2.464 (67)*

Tabella 2: Valori formantici medi delle vocali toniche e deviazione standard (tra parentesi) separati per il fattore 'accento', con i valori di t, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

La durata è il parametro decisamente più sensibile: le vocali sono significativamente più lunghe se colpite da accento di frase (solo per [e] la differenza non raggiunge la significatività statistica). L'incremento percentuale va da un minimo di 17% per la vocale [u] a un massimo di 39% per [ɛ]. La sensibilità della prima formante è marcata sull'asse anteriore e poco rilevante sull'asse posteriore: sotto accento di frase, la prima formante raggiunge valori significativamente più elevati per [a] e per [ɛ] (per [ɔ] siamo prossimi alla significatività) e valori significativamente più bassi per le vocali medie (e anche per le alte, ma senza raggiungere la significatività statistica). Gli effetti sulla seconda formante sono più deboli: sotto accento frasale, F2 diminuisce per [a] e per [o], mentre aumenta per [ɛ] e [i]. Sotto accento le vocali medie (e, tendenzialmente, anche le medio-alte) diventano dunque più tese.

Per valutare l'effetto della posizione della vocale tonica, nell'analisi statistica sono considerate tutte e tre le opzioni sebbene i dati non siano perfettamente bilanciati. Nel caso in cui l'ANOVA a una via abbia rilevato la significatività statistica, una serie di test *post-hoc* (Sidak) ha valutato quali delle tre posizioni fossero significativamente differenti (le differenze sono indicate dal segno di disuguaglianza nella tabella 3). Ancora una volta, è la durata ad essere maggiormente colpita dalle variazioni nella posizione della vocale tonica: ad eccezione di [u] tutti i confronti sono risultati significativi.[14] La vocale tonica in posizione finale di enunciato o di turno è significativamente più lunga delle vocali in posizione interna e iniziale di parola; sono invece scarsamente differenziate le durate in posizione iniziale e interna.

[14] La vocale [u] tra l'altro non mostra effetti significativi su nessuno dei tre parametri misurati; per ragioni di spazio viene omessa.

posizione	a	ɛ	e	i	ɔ	o
p.iniziale	109 (36)	87 (27)	65 (44)	91 (29)	100 (24)	80 (29)
p.interna	97 (37)	103 (49)	69 (28)	73 (30)	91 (63)	84 (39)
p.finale	148 (28)	132 (37)	120 (39)	111 (24)	141 (35)	115 (38)
ANOVA	F=34.810 (2,211)*** p.fin.≠p.iniz.; p.fin.≠p.int.	F=9.717 (2,163)*** p.fin.≠p.iniz.; p.fin.≠p.int.	F=12.764 (2,74)*** p.fin.≠p.iniz.; p.fin.≠p.int.	F=29.401 (2,150)*** p.fin.≠p.iniz.; p.fin.≠p.int.; p.int.≠p.iniz.	F=9.857 (2,97)*** p.fin.≠p.int	F=4.582 (2,72)* p.fin.≠p.iniz. p.fin.≠p.int.

Tabella 3: Valori temporali delle vocali toniche e deviazione standard (tra parentesi) separati per il fattore 'posizione', con i valori di F, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

A decrescere, si registra qualche effetto – vieppiù debole – della posizione sulla seconda formante su [ɛ] e [e], con una tendenza ad occupare spazi periferici nelle vocali toniche finali di parola rispetto a quelle iniziali e interne. La stessa tendenza, *mutatis mutandis*, si registra in maniera più robusta sull'asse posteriore per [o].[15] Un effetto più debole è presente sulla prima formante di [ɔ]: in posizione iniziale il valor medio è 606 Hz (d.s. 62), in posizione interna è 556 Hz (d.s. 74), in posizione finale è 593 (d.s. 90).[16]

Gli effetti della posizione sulla durata sono visibili anche nel sistema atono (v. tab. 4), ove – lo ricordiamo – il confronto è solo tra vocali atone in posizione interne di parola e vocali atone in posizione finale di parola ma non in posizione finale di enunciato (dunque le cosiddette 'finali false'):

[15] Effetti più deboli su F2 di [ɛ]: in posizione iniziale 1653 Hz (d.s. 150), in posizione interna 1721 Hz (d.s. 136), in posizione finale 1752 Hz (d.s. 155): F = 3.503 (2,163)* (pos. finale ≠ pos. iniziale); su F2 di [e]: in posizione iniziale 1859 Hz (d.s. 165), in posizione interna 1956 Hz (d.s. 145), in posizione finale 2034 Hz (d.s. 122): F = 4.766 (2,74)* (pos. finale ≠ pos. iniziale). L'effetto su F2 appare più robusto nel caso di [o]: in posizione iniziale 1269 Hz (d.s. 257), in posizione interna 1107 Hz (d.s. 178), in posizione finale 903 Hz (d.s. 145): F = 14.454 (2,72)*** (pos. finale ≠ pos. iniziale; pos. finale ≠ pos. interna, pos. interna ≠ pos. iniziale).

[16] F = 3.346 (2, 97)*.

v	*F1*			*F2*			*durata*		
	interne di parola	finali di parola	valori di t	interne di parola	finali di parola	valori di t	interne di parola	finali di parola	valori di t
a	608 (101)	571 (93)	2.531 (180)*	1399 (186)	1445 (136)	n.s.	63 (18)	73 (38)	-2.309 (180)*
e	412 (47)	441 (83)	-2.415 (132)*	1777 (185)	1773 (182)	n.s.	52 (15)	62 (33)	-2.067 (132)*
i	370 (85)	377 (64)	n.s.	1999 (199)	1977 (187)	n.s.	57 (59)	67 (37)	n.s.
o	453 (90)	474 (84)	n.s.	1286 (247)	1238 (199)	n.s.	51 (16)	62 (35)	-2.473 (141)*

Tabella 4: Valori formantici medi delle vocali atone e deviazione standard (tra parentesi) separati per il fattore 'posizione', con i valori di t, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

Le vocali atone in posizione finale di parola tendono ad essere più lunghe rispetto a quelle in posizione interna; per la prima formante mostrano la tendenza a raggiungere una posizione centrale dello spazio vocalico (cfr., per [a], Vayra, 1991). I movimenti appaiono irrilevanti per la seconda formante.

Per quanto concerne il fattore semantico (parole piene *vs.* parole vuote), non per tutte le vocali è stato possibile raggiungere una numerosità adeguata:[17] in particolare, mancano dati per il confronto relativo alle vocali [a] e [i] (v. tab. 5).

v	*F1*			*F2*			*durata*		
	parole vuote	parole piene	valori di t	parole vuote	parole piene	valori di t	parole vuote	parole piene	valori di t
ɛ	510 (86)	575 (81)	-4.453 (169) ***	1685 (155)	1734 (139)	n.s.	73 (31)	117 (45)	-7.148 (169) ***
e	427 (53)	402 (58)	n.s.	1837 (106)	1980 (149)	-3.487 (76)**	57 (18)	80 (40)	-3.445 (76)**
ɔ	569 (76)	576 (82)	n.s.	1149 (134)	1098 (138)	n.s.	102 (82)	111 (45)	n.s.
o	477 (65)	446 (68)	n.s.	1227 (189)	1038 (218)	3.753 (74) ***	76 (31)	95 (40)	-2.080 (74)*
u	431 (59)	381 (70)	2.500 (67)*	986 (233)	934 (182)	n.s.	85 (39)	86 (37)	n.s.

Tabella 5: Valori formantici medi delle vocali toniche e deviazione standard (tra parentesi) separati su base semantica, con i valori di t, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

[17] Il campione risulta sbilanciato a favore delle parole piene. Il confronto statistico è stato ritenuto valido quando, per ciascuna vocale, le parole vuote hanno raggiunto la soglia di 15 entrate.

Ad ogni buon conto, rileviamo come le vocali toniche di parole lessicalmente piene siano nel complesso più lunghe. Per quanto concerne i valori spettrali, la vocale [ɛ] presenta un valore significativamente più elevato della prima formante, mentre [u] appare più periferica. Nel caso delle vocali medie, la seconda formante assume valori significativamente più elevati per [e] e significativamente più bassi per [o].

Un quadro diverso è offerto dal vocalismo atono, come mostra la tabella 6:

v	*F1*			*F2*			*durata*		
	parole vuote	parole piene	valori di t	parole vuote	parole piene e	valori di t	parole vuote	parole piene	valori di t
a	598 (87)	586 (103)	n.s.	1359 (122)	1450 (164)	-4.229 (191) ***	64 (33)	73 (32)	n.s.
e	412 (51)	432 (75)	n.s.	1712 (181)	1784 (181)	n.s.	47 (13)	65 (40)	104.695 (139) ***
i	346 (69)	377 (79)	n.s.	1896 (122)	2005 (196)	-2.116 (170)*	55 (10)	65 (60)	n.s.
o	439 (64)	472 (98)	n.s.	1281 (166)	1250 (231)	n.s.	48 (28)	60 (30)	n.s.

Tabella 6: Valori formantici medi delle vocali atone e deviazione standard (tra parentesi) separati su base semantica, con i valori di t, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

Le differenze, individuabili spesso soltanto come mere tendenze nel confronto tra medie,[18] sono in pochissimi casi significative: rispetto al vocalismo tonico, il sistema atono è scarsamente differenziato sulla base di fattori semantici.

4.2 Vocali toniche e atone toscane

Il confronto fra sistema tonico e sistema atono è viziato da una difficoltà procedurale relativa all'incerto statuto delle vocali atone medie (cfr. Mioni, 1993: 122).[19] Tuttavia, da un altro punto di vista, diventa particolarmente robusto poiché è un confronto riferito ai medesimi locutori che nella medesima condizione sperimentale producono sia vocali toniche che vocali atone. Le principali differenze tra sistema atono e sistema tonico nel complesso delle vocali 'toscane' sono riassunte nella tabella 7:

[18] Nel caso della durata, tutte e quattro le vocali sono più lunghe nelle parole piene.

[19] Anche alla luce di considerazioni tipologiche (nella riduzione vocalica di tipo fonologico le vocali medie molto spesso si innalzano) il confronto statistico (t test) è stato fatto, rispettivamente, tra le atone [e] e [o] e le due vocali toniche medio-alte [e] e [o].

v	*F1*		valori di t	*F2*		valori di t	*durata*		valori di t
	vocali atone	vocali toniche		vocali atone	vocali toniche		vocali atone	vocali toniche	
a	590 (98)	674 (108)	-8.179 (406) ***	1424 (158)	1336 (128)	6.161 (406) ***	71 (33)	108 (41)	-10.202 (406) ***
e	429 (72)	407 (57)	2.328 (218)*	1773 (182)	1956 (154)	-7.540 (218)***	62 (38)	76 (38)	-2.671 (218)**
i	374 (78)	350 (68)	3.007 (323) **	1996 (193)	2154 (220)	-6.898 (323) ***	58 (58)	90 (33)	-4.955 (323) ***
o	468 (95)	456 (68)	n.s.	1254 (224)	1102 (226)	4.789 (227) ***	59 (30)	89 (38)	-6.473 (227) ***

Tabella 7: Valori formantici medi delle vocali atone e deviazione standard (tra parentesi) separati su base semantica, con i valori di t, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

Le principali differenze si collocano sul dominio temporale (vocali toniche sempre significativamente più lunghe rispetto a quelle atone) e sulla seconda formante (vocali toniche sempre più periferiche delle corrispondenti atone). Solo nel caso della vocale bassa si raggiunge una elevata significatività statistica per la prima formante, che ha valori molto più elevati sotto accento. Per le vocali [e i o] si registra una tendenza ad assumere posizioni meno basse e quindi meno aperte nel vocalismo tonico. Nel parlato semispontaneo il vocalismo atono non è più variabile del vocalismo tonico: le deviazioni standard sono in entrambe i raggruppamenti relativamente elevate.

4.3 Le vocali nello spazio

Almeno negli stili controllati i tre punti di rilevamento presentano per il vocalismo tonico differenze sostanziali: è noto infatti come l'area occidentale della Toscana mostri consistenti fenomeni di abbassamento per le medio-basse, articolazioni più posteriori per [a] (Giannelli, 2000: 63; Calamai, 2004), insieme a una complessa, e solo parzialmente studiata, interazione tra frequenze formantiche, andamenti di f0, durata (Marotta, Calamai & Sardelli, 2004). Il vocalismo elicitato in stili più liberi appare senza dubbio più variabile (le deviazioni standard, specie per la seconda formante, sono molto elevate) e pertanto più difficilmente trattabile a un confronto statistico. Le aree di esistenza al 68% riportate nelle Figure 1-3 evidenziano anche ad occhio nudo questa estrema variabilità: [20]

[20] In rosso sono rappresentate le ricorrenze di [a], in verde chiaro quelle di /ɛ/, in arancione quelle di [e], in verde acqua sono i dati per [i]; per l'asse posteriore i colori scelti sono i seguenti: verde ocra per [ɔ], blu per [o] e verde per [u].

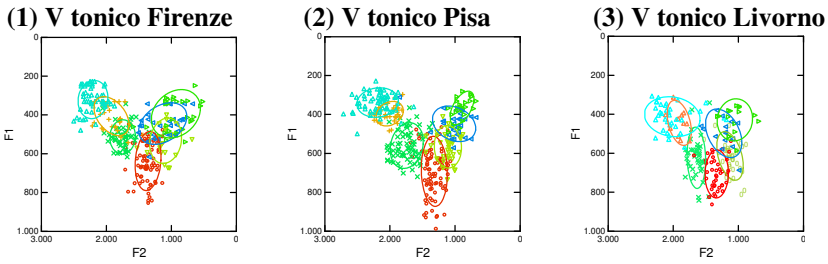


Figure 1-3: Ellissi di dispersione al 68% dei vocalismi tonici di Firenze, Pisa e Livorno

A Firenze (Fig. 1), la vocale [i] presenta l'area più compatta; sul fronte posteriore c'è una consistente sovrapposizione tra [o] e [u]. A Pisa (Fig. 2), si assiste a un abbassamento marcato di quasi tutte le ellissi e a un consistente avvicinamento di [a] a [ɛ], realizzato spesso come [æ]. Rispetto a Livorno (cfr. Fig. 3), viene mantenuta ancora una ragionevole distinzione tra [e] e [i], mentre sull'asse posteriore è da notare una sovrapposizione tra [o] e [u]: la medio-alta, in particolare, presenta diversi esiti anteriorizzati, foneticamente realizzati come [ɵ]. A Livorno le aree vocaliche mostrano una minore distinzione reciproca e una maggiore sovrapposizione: in particolare, le medio-basse si avvicinano in maniera consistente all'area di [a], già molto posteriore. Si assiste inoltre a una sovrapposizione marcata delle ellissi di [e] e [i]: la vocale anteriore alta nel parlato spontaneo livornese è una vocale rilassata con valori di F2 che scendono fino a 1600 Hz (cfr. § 4). A tale proposito, presentiamo lo spettrogramma della parola *baracchina* (cfr. Figura 4) prodotta dal soggetto livornese GS in cui è visibile una realizzazione di [i] rilassata con valori di F1 di 490 Hz e F2 di 1587 Hz.

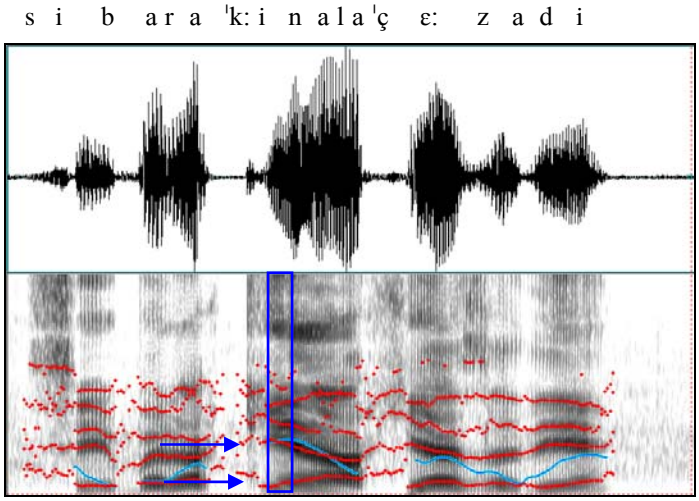


Figura 4: Forma d'onda e spettrogramma del sintagma *sì baracchina la chiesa di?* realizzato dal soggetto GS-M-LI {audio 1}

A ben guardare in tutte e tre le località, l'articolazione di [o] si mostra relativamente anteriorizzata sull'asse posteriore; resta da ancora da verificare tuttavia se si tratti di un tratto peculiare del vocalismo toscano *tourt-court* o se siamo di fronte ad un fenomeno di semplice *undershoot*.

Le figure 5-6 presentano un'esemplificazione di questo fenomeno, in cui le formanti di /o/ per il parlante livornese misurano rispettivamente 417 Hz (F1) e 1517 Hz (F2), mentre per il locutore fiorentino i valori sono di 524 Hz e 1319 Hz.

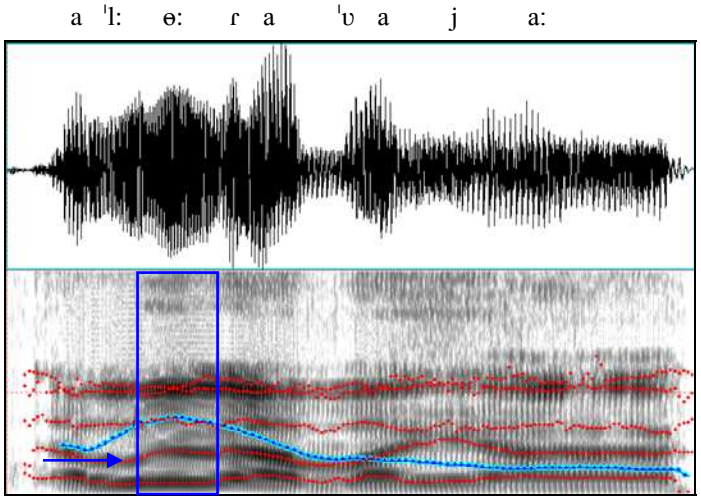


Figura 5: Forma d'onda e spettrogramma del sintagma *allora vai (al quartiere San Jacopo)* realizzato dal soggetto GS-M-LI {audio 2}

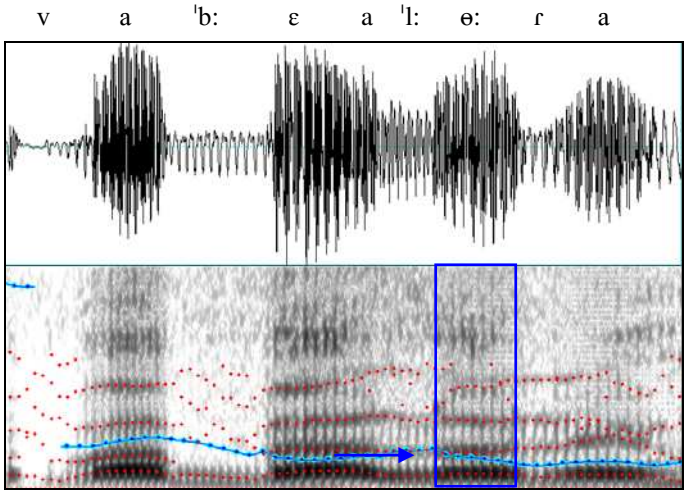


Figura 6: Forma d'onda e spettrogramma del sintagma *vabbè allora (i cuochi)* realizzato dal soggetto SdG-M-FI {audio 3}

Laddove possibile, per il confronto inter-località è stata calcolata una ANOVA a una via, mentre una serie di test *post-hoc* (Sidak) hanno permesso di valutare quali delle tre località fossero significativamente differenti (le differenze tra i punti sono riportate nell'ultimo rigo delle tabelle con il segno di disuguaglianza). Le tabelle 8 e 9 riportano i risultati relativi alle prime due formanti:

luogo	a	ɛ	e	i	ɔ	o	u
Firenze	637 (99)	513 (57)	407 (64)	321 (64)	533 (74)	445 (68)	394 (79)
Pisa	692 (117)	560 (76)	396 (40)	338 (49)	566 (70)	448 (58)	366 (54)
Livorno	706 (79)	623 (101)	444 (68)	412 (65)	614 (79)	498 (76)	429 (66)
ANOVA	F=8.409 (2,212)*** Fi ≠ Li; Fi ≠ Pi; Li = Pi	F=23.342 (2,168)*** Fi ≠ Li; Fi ≠ Pi; Li ≠ Pi	F=3.136 (2,76)* Fi = Li; Fi = Pi; Li ≠ Pi	F=28.378 (2,150)*** Fi ≠ Li; Fi = Pi; Li ≠ Pi	F=9.074 (2,98)*** Fi ≠ Li; Fi = Pi; Li ≠ Pi	F=3.685 (2,73)* Fi ≠ Li; Fi = Pi; Li ≠ Pi	F=4.142 (2,66)* Fi = Li; Fi = Pi; Li ≠ Pi

Tabella 8: Valori medi della prima formante (sistema tonico) e deviazione standard (tra parentesi) separati per il fattore luogo, con i valori di F, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

luogo	a	ɛ	e	i	ɔ	o	u
Fi	1357 (130)	1749 (127)	1922 (173)	2205 (149)	1088 (160)	1132 (235)	931 (239)
Pi	1318 (130)	1742 (166)	2004 (134)	2188 (207)	1109 (134)	1014 (211)	893 (115)
Li	1334 (115)	1646 (91)	1909 (109)	2015 (272)	1129 (130)	1226 (170)	1054 (166)
ANOVA	n.s.	7.876 (2,168)** Fi ≠ Li; Fi = Pi; Li ≠ Pi	3.186 (2,76)* Fi = Li; Fi = Pi; Li = Pi	10.588 (2,150)*** Fi ≠ Li; Fi = Pi; Li ≠ Pi	n.s.	5.419 (2,73)** Fi = Li; Fi = Pi; Li ≠ Pi	3.803 (2,66)* Fi = Li; Fi = Pi; Li ≠ Pi

Tabella 9: Valori medi della seconda formante (sistema tonico) e deviazione standard (tra parentesi) separati per il fattore luogo, con i valori di F, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05).

Gli effetti del luogo sono robusti sulla prima formante, meno forti sulla seconda, ancor meno forti sulla durata – dove gli effetti sono debolmente significativi solo su /a/ e su /ɛ/.[21] Le scarse significatività sul parametro della durata rappresentano a nostro avviso anche una prova indiretta del fatto che i dialoghi sono stati ottenuti a una velocità d'eloquio relativamente omogenea.

[21] Riportiamo in nota i valori medi della durata e le deviazione standard (tra parentesi) separati per il fattore 'luogo', con i valori di F, i gradi di libertà e il livello della significatività statistica per i due confronti significativi (*** p <.001, ** p<.01, *.p<.05). Vocale [a]: Fi 99 (36), Li 116 (41), Pi 113 (43); F =3.436 (2,212)*; Fi = Li; Fi = Pi; Li = Pi; vocale [ɛ]: Fi 93 (31), Li 117 (54), Pi 112 (47); F =4.216 (2,168)* Fi ≠ Li; Fi = Pi; Li = Pi. Ci sembra comunque rilevante che la vocale 'bandiera' di Livorno (e di Pisa) sia significativamente più lunga rispetto a quella di Firenze.

Il confronto statistico fra le tre località mostra come la coppia con il numero più elevato di differenze significative (e pertanto con un vocalismo sensibilmente diverso, o – se si preferisce – meno simile) sia quella costituita da Pisa e Livorno: il risultato è, per certi versi, inaspettato se si pensa che la letteratura dialettologica etichetta come sistema 'pisano-livornese' il dialetto parlato nelle due città (cfr. Giannelli, 2000). Soltanto la vocale [a] non presenta differenziazioni (confermando dunque una generale impressione di arretramento di [a] lungo la fascia occidentale toscana, rilevata in più sedi da Luciano Giannelli e Fabrizio Franceschini), mentre le vocali medie (medio-alte e medio-basse) così come le vocali alte sono significativamente diverse. Dalle vocali livornesi sono molto differenti anche quelle fiorentine, mentre i due sistemi più simili (o, se si preferisce, meno diversi) sono il fiorentino e il pisano. Il comportamento della seconda formante conferma, in maniera più sbiadita, quanto rilevato per la prima: Pisa e Firenze sono sostanzialmente indifferenziati, mentre Livorno è diverso da Pisa per quattro vocali che a Livorno appaiono, sull'asse anteriore, meno anteriori e, sull'asse posteriore, meno posteriori, producendo in sostanza un restringimento dello spazio vocalico.

Anche nel vocalismo atono il fattore 'luogo' mostra di avere un peso, come evidenziano i tre grafici con le ellissi di dispersione (v. Figg. 7-9) e come prova l'analisi statistica, soprattutto per quanto concerne la prima formante (cfr. tabelle 10-11).

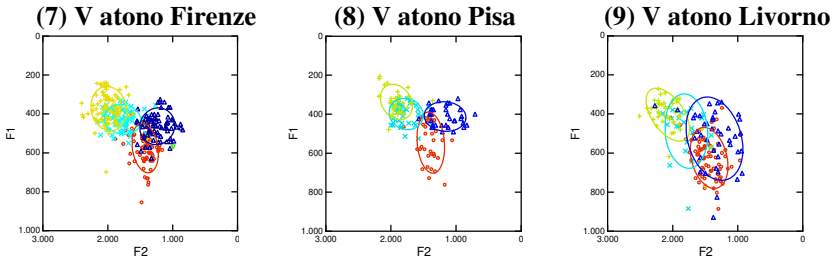


Figure 7-9: Ellissi di dispersione al 68% dei vocalismi atoni di Firenze, Pisa e Livorno.

Colpiscono subito il maggiore disordine e le sovrapposizioni estreme presenti nel vocalismo livornese, caratterizzato da un consistente abbassamento, evidenziato anche dai valori numeri riportati nella tabella 10.

luogo	a	e	i	o
Fi	575 (83)	428 (50)	373 (74)	463 (61)
Pi	546 (100)	399 (52)	337 (58)	412 (48)
Li	625 (101)	488 (119)	404 (87)	527 (137)
ANOVA	10.323 (2,190)*** Fi ≠ Li; Fi = Pi; Li ≠ Pi	13.197 (2,138)*** Fi ≠ Li; Fi ≠ Pi; Li ≠ Pi	7.727 (2,169)** Fi ≠ Li; Fi ≠ Pi; Li ≠ Pi	17.216 (2,150)*** Fi ≠ Li; Fi ≠ Pi; Li ≠ Pi

Tabella 10: Valori medi della prima formante (sistema atono) e deviazione standard (tra parentesi) separati per il fattore luogo, con i valori di F, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

Le vocali atone di Livorno hanno sempre la prima formante significativamente più bassa; mentre i dati relativi alla seconda mostrano nel complesso una minore diversificazione:

luogo	a	e	i	o
Fi	1425 (133)	1771 (184)	1985 (179)	1239 (176)
Pi	1388 (135)	1767 (168)	1916 (157)	1168 (204)
Li	1439 (188)	1789 (207)	2074 (216)	1357 (280)
ANOVA	n.s.	n.s.	7.426 (2,169)** Fi ≠ Li; Fi = Pi; Li ≠ Pi	7.728 (2,150)** Fi ≠ Li; Fi = Pi; Li ≠ Pi

Tabella 11: Valori medi della seconda formante (sistema atono) e deviazione standard (tra parentesi) separati per il fattore luogo, con i valori di F, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

Le vocali livornesi si differenziano nella stessa misura dalle vocali fiorentine e da quelle pisane, mentre le differenze tra vocali pisane e vocali fiorentine sono nel complesso ridotte. A Livorno, le vocali sono abbassate anche nel sistema atono. Sempre per il vocalismo livornese, si registra una maggiore anteriorizzazione delle vocali [i] e [o|: ci limitiamo a registrare il dato, non potendo stabilire al momento se sia dovuto a fattori geolinguistici o meramente contestuali. Nel complesso, la differenza fra le tre località è decisamente minore nella durata, ove la significatività è presente solo per [i].[22] Dal punto di vista temporale, dunque, il sistema atono tende a comportarsi come quello tonico, mostrando nelle tre località tendenze omogenee per quanto concerne la velocità d'eloquio.

4.4 Il fattore 'accento' nelle tre località

L'analisi separata delle tre località per valutare gli effetti del fattore 'accento' mostra come questo abbia, nella sostanza, effetti simili nei tre punti per il parametro durata ma effetti in parte divergenti sul piano timbrico (cfr. tabelle 12-14).

[22]Vocale[i]: Fi 56 (22), Li 95 (100), Pi (46) F=10.407(2,169)*** Fi ≠ Li; Fi = Pi; Li ≠ Pi.

Vocali toniche fiorentine									
v	*F1*			*F2*			*durata*		
	accento lessicale	accento frasale	valori di t	accento lessicale	accento frasale	valori di t	accento lessicale	accento lessicale	valori di t
a	617 (93)	677 (99)	-2.709 (1,81)**	1366 (139)	1340 (111)	n.s.	86 (28)	126 (36)	-5.597 (1,81)***
ɛ	513 (60)	515 (54)	n.s.	1711 (112)	1809 (129)	-3.013 (1,54)**	83 (31)	106 (29)	-2.863 (1,54)**
e	437 (40)	377 (61)	3.461 (1,32)**	1917 (175)	1939 (169)	n.s.	65 (25)	76 (27)	n.s.
i	312 (55)	323 (66)	n.s.	2144 (160)	2223 (142)	n.s.	78 (35)	99 (27)	-2.222 (1,52)*
ɔ	555 (81)	511 (62)	n.s.	1090 (203)	1087 (97)	n.s.	99 (31)	94 (33)	n.s.
o	462 (62)	404 (67)	2.298 (1,28)*	1169 (223)	1043 (253)	n.s.	80 (25)	106 (25)	-2.651 (1,28)*
u	417 (75)	326 (48)	2.992 (1,26)**	968 (243)	819 (203)	n.s.	78 (32)	100 (19)	n.s.

Tabella 12: Valori formantici medi separati per il fattore accento, con i valori di t, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

A Firenze la durata appare sensibile all'accento per quattro vocali su sette (con le vocali più lunghe se sotto accento di frase); anche la prima formante appare sensibile ma con effetti differenti: i valori aumentano per [a], mentre diminuiscono per le vocali medie e le alte, che sotto accento diventano più tese.

Vocali toniche pisane									
v	*F1*			*F2*			*durata*		
	accento lessicale	accento frasale	valori di t	accento lessicale	accento frasale	valori di t	accento lessicale	accento lessicale	valori di t
a	657 (111)	745 (107)	-3.784 (89)***	1344 (137)	1279 (111)	2.378 (89)*	96 (36)	139 (42)	-5.228 (89)***
ɛ	534 (78)	597 (55)	-3.852 (73)***	1709 (171)	1788 (148)	-2.081 (73)*	80 (30)	156 (30)	-10.796 (73)***
e	395 (44)	396 (37)	n.s.	1993 (141)	2017 (130)	n.s.	63 (27)	92 (47)	-2.193 (32)*
i	337 (45)	340 (53)	n.s.	2136 (143)	2230 (241)	n.s.	72 (30)	107 (27)	-4.846 (61)***
ɔ	537 (56)	605 (69)	-3.456 (38)**	1119 (154)	1096 (105)	n.s.	82 (38)	147 (42)	-4.976 (38)***
o	449 (71)	446 (31)	n.s.	1047 (240)	962 (150)	n.s.	64 (30)	117 (35)	-4.326 (29)***
u	368 (67)	364 (29)	n.s.	884 (114)	906 (114)	n.s.	80 45)	87 (39)	n.s.

Tabella 13: Valori formantici medi separati per il fattore accento, con i valori di t, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

Anche a Pisa la durata appare sempre maggiore sotto accento di frase (l'unica vocale per cui non viene raggiunta la significatività statistica è [u]). Sotto accento di frase, la prima formante diventa più elevata per [a] e le vocali medio-basse; la seconda formante diminuisce per /a/ e aumenta nelle vocali anteriori.

Vocali toniche livornesi									
v	*F1*			*F2*			*durata*		
	accento lessicale	accento frasale	valori di t	accento lessicale	accento frasale	valori di t	accento lessicale	accento lessicale	valori di
a	700 (78)	720 (83)	n.s.	1346 (116)	1304 (112)	n.s.	107 (39)	138 (40)	-2.265 (39)*
ɛ	588 (107)	662 (79)	-2.474 (38)*	1629 (96)	1665 (84)	n.s.	94 (49)	142 (51)	-3.004 (38)**
i	400 (63)	427 (67)	n.s.	2025 (198)	2004 (344)	n.s.	63 (30)	106 (26)	-4.618 (34)***
ɔ	590 (69)	673 (75)	-3.011 (33)**	1164 (128)	1043 (92)	3.095 (33)*	113 (82)	128 (47)	n.s.
u	412 (59)	445 (73)	n.s.	1067 (166)	1041 (177)	n.s.	73 (17)	115 (48)	-2.336 (14)*

Tabella 14: Valori formantici medi separati per il fattore accento, con i valori di t, i gradi di libertà e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

Per scarsità di ricorrenze non è stato possibile confrontare i valori relativi alle vocali medio-alte. Ad ogni buon conto, sotto accento di frase, la durata appare sempre maggiore; la prima formante è più elevata per le vocali medio-basse, mentre la seconda formante assume valori inferiori per la vocale [ɔ].

In sintesi, è comune alle tre località una maggiore sensibilità da parte della prima formante alle variazioni accentuali rispetto alla seconda. Tuttavia, non tutte le vocali sono sensibili allo stesso modo:[23] le vocali bandiera della Toscana occidentale – le medio-basse [ɛ] e, in misura minore, [ɔ] – mostrano di essere sensibili alla variazione accentuale proprio a Pisa e a Livorno, ma non a Firenze, dove restano sostanzialmente stabili (vedi sul piano qualitativo § 5). Infine, anche se tutte e tre le località mostrano una elevata sensibilità del parametro durata ai cambiamenti relativi all'accento, appare diversa l'entità dell'incremento: laddove il confronto è significativo, nelle vocali fiorentine l'incremento percentuale ha come tetto massimo il 32% (per [a]), mentre nelle vocali pisane raggiunge (peraltro proprio per [ɛ]) il 49%.[24]

4.5 Sistemi tonici e sistemi atoni nelle tre località

I risultati del confronto tra sistema atono e sistema tonico rivestono un interesse particolare proprio in merito al tema della riduzione perché potrebbero da un lato mostrare tendenze identiche (la riduzione avviene nello stesso modo nelle tre diverse località), ovvero mostrare comportamenti difformi, imputabili a condizionamenti geolinguistici. Le tabelle 15-17 mostrano i dati suddivisi per località:

[23] Il ragionamento vale soprattutto per le località di Pisa e Firenze, essendo parziali i dati di Livorno.

[24] A Livorno, l'incremento massimo è raggiunto da [i] (41%).

Vocali fiorentine									
v	*F1*			*F2*			*Durata*		
	tonica	atona	valori di t	tonica	atona	valori di t	tonica	atona	valori di t
a	637 (99)	575 (83)	-4.358 (161)***	1358 (130)	1425 (133)	3.285 (161)**	99 (36)	64 (27)	-7.008 (161)***
e	407 (64)	427 (50)	n.s.	1922 (173)	1771 (184)	-4.063 (109)***	71 (26)	57 (25)	-2.780 (109)**
i	321 (64)	373 (74)	4.285 (143)***	2206 (149)	1985 (179)	-7.619 (143)***	95 (30)	56 (22)	-9.028 (143)***
o	445 (68)	463 (61)	n.s.	1132 (235)	1239 (176)	2.546 (104)*	88 (28)	57 (25)	-5.637 (104)***

Tabella 15: Valori formantici medi per il sistema tonico e per quello atono, con deviazione standard (tra parentesi), i valori di t, i gradi di libertà (tra parentesi) e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05).

Nel campione fiorentino (tab. 15) la durata è sempre significativa, con le vocali toniche sensibilmente più lunghe della atone. La prima formante è maggiore nelle toniche per [a], mentre è minore per [i]; la seconda formante è maggiore per [e] e [i] mentre è minore per [a] e [o]. Le vocali medie mantengono posizioni simili per quanto riguarda F1; la /i/ tonica appare più tesa della corrispondente atona.

Vocali pisane									
v	*F1*			*F2*			*Durata*		
	tonica	atona	valori di t	tonica	atona	valori di t	tonica	atona	valori di t
a	692 (117)	546 (100)	6.562 (125)***	1318 (131)	1388 (135)	-2.695 (125)**	113 (43)	73 (39)	5.016 (125)***
e	396 (40)	399 (52)	n.s.	2004 (134)	1767 (168)	-6.648 (73)***	77 (40)	64 (53)	n.s.
i	341 (52)	337 (58)	n.s.	2185 (207)	1916 (157)	6.616 (96)***	91 (33)	46 (11)	7.665 (96)***
o	448 (58)	412 (49)	-2.758 (65)**	1014 (211)	1169 (205)	3.028 (65)**	85 (41)	63 (37)	-2.305 (65)*

Tabella 16: Valori formantici medi per il sistema tonico e per quello atono, con deviazione standard (tra parentesi), i valori di t, i gradi di libertà (tra parentesi) e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

Nel campione pisano (tab. 16) le differenze di durata sono un po' meno significative rispetto al campione fiorentino, anche se le vocali toniche sono generalmente sempre più lunghe delle corrispondenti atone. La prima formante è maggiore nelle toniche per [a]; si mantiene stabile sulle vocali anteriori e più elevata per [o]. La seconda formante è maggiore per le toniche [e] e [i], è minore per [a] e [o].

Vocali livornesi									
v	*F1*			*F2*			*Durata*		
	tonica	atona	valori di t	tonica	atona	valori di t	tonica	atona	valori di t
a	705 (79)	625 (102)	-4.304 (115)***	1333 (115)	1439 (188)	3.246 (115)**	116 (41)	76 (35)	-5.561 (116)***
e	444 (68)	488 (119)	n.s.	1909 (109)	1789 (207)	n.s.	90 (60)	75 (41)	n.s.
i	412 (65)	404 (87)	n.s.	2015 (273)	2074 (216)	n.s.	83 (35)	95 (100)	n.s.
o	515 (94)	527 (137)	n.s.	1220 (174)	1357 (280)	n.s.	94 (50)	59 (31)	-3.201 (56)**

Tabella 17: Valori formantici medi per il sistema tonico e per quello atono, con deviazione standard (tra parentesi), i valori di t, i gradi di libertà (tra parentesi) e il livello della significatività statistica (*** p <.001, ** p<.01, * p<.05)

A Livorno i due sistemi appaiono più simili. Solo in due casi, le vocali toniche sono significativamente più lunghe delle corrispondenti atone. Soltanto la vocale [a] presenta differenze nelle due formanti

In sintesi, nel passaggio da sistema tonico a quello atono le maggiori differenze sono rintracciabili sul piano temporale (v. Farnetani & Busà, 1999). Una differenza legata alla variabile spazio è rintracciabile nella consistenza dei cambiamenti, massimi a Firenze, intermedi a Pisa, limitati a Livorno, il cui sistema atono sembra essere nel complesso più simile a quello tonico. Colpisce, nelle tre località, il comportamento omogeneo di [i] che non mostra tendenze verso una centralizzazione nel passaggio da tonico a atono: al contrario, a Firenze, presenta una maggiore chiusura, a Pisa e Livorno una sostanziale stabilità sulla prima formante, a Firenze e Pisa una tendenza verso una posizione più posteriore. In mancanza di dati sulla corrispondente vocale alta posteriore non ci sembra opportuno avanzare osservazioni più generali relative alle tendenze presenti nel passaggio da sistemi tonici a quelli atoni. Fuor di dubbio è la relativa stabilità di [i] che da un punto di vista di fonetica generale mostra ancora una volta una maggiore resistenza alla coarticolazione, se confrontata con le altre vocali (si osservino a questo proposito, i valori della deviazione standard della prima formante di [i] nelle tre località).

Per valutare se lo spazio vocalico è sempre più ampio nel vocalismo tonico rispetto al corrispondente atono è stata calcolata l'area del poligono formato dalle linee che congiungono le vocali rappresentate su un piano con le coordinate F1/F2:[25] in tutte e tre le località il sistema atono occupa uno spazio minore rispetto al corrispondente tonico. Appaiono leggermente diverse le percentuali di 'restringimento': il vocalismo pisano è quello che si modifica di più (da 151203 a 47099 Hz², con una riduzione del 69%), seguito da quello fiorentino (da 108708 a 43372 Hz², con una riduzione del 60%), seguito infine da

[25] Il calcolo dell'area del poligono fornisce un'indicazione sull'eventuale contrazione dello spazio vocalico nei casi di tendenza alla centralizzazione (cfr. Pätzold & Simpson, 1997: 230); non è tuttavia in grado di identificare movimenti e dinamiche della contrazione in riferimento alle singole vocali.

quello livornese il quale, occupando uno spazio minore già nel sistema tonico (77107 Hz²), si riduce del 54% (35353 Hz²).

5. ANALISI QUALITATIVA: VOCALI COME SENTIERI

In questa ultima parte della nostra ricerca presentiamo i risultati relativi ad una indagine di tipo qualitativo – che ci proponiamo di approfondire in futuro – dedicata all'ispezione fine dell'evoluzione temporale dei tracciati formantici. Intendiamo verificare se è possibile rendere conto di alcune differenze geolinguistiche in base alle differenze nelle traiettorie formantiche. È ben noto che l'andamento temporale delle formanti non è mai piatto, neppure in quelle porzioni di segmento considerate stazionarie, che mostrano al contrario una sorta di dinamicità interna, tanto da parlare di *Vowel-Inherent Spectral Change*. Sempre più spesso quindi, le vocali sono considerate come segmenti dinamici (Lisker, 1984; Di Benedetto, 1989; Cerrato & Cutugno, 1994; Hillenbrand *et al.*, 1995) e non come punti statici.

Alla luce di queste considerazioni, presentiamo i risultati del confronto diatopico dei tracciati delle ricorrenze relative alla parola *destra* nelle tre località indagate (cinque per punto): si tratta quindi di materiale segmentale identico, caratterizzato da una estrema omogeneità di contesto sia dal punto di vista prosodico che pragmatico (*sulla destra*, *vai a destra*, *devi andare a destra*…)[26].

Per queste quindici entrate, i valori analizzati con *Praat* sono stati estratti mediante uno *script* realizzato da Beat Siebenhaar (Università di Lipsia) che permette di misurare i valori formantici e di durata dal 5% fino al 95% della durata di tutti i segmenti etichettati. Lo *script* si basa sull'etichettatura manuale dei foni tramite la creazione di un file *TextGrid* in cui vengono salvate tutte le segmentazioni (cfr. Figura 10). Dopo l'applicazione dello *script* viene creato automaticamente un file *Log.xls* contenente i valori dei dati misurati. In questo modo, una volta etichettato il *corpus* con accuratezza, sarà possibile utilizzare il materiale sonoro per altri tipi di analisi strumentale. Per l'etichettatura ci si è avvalsi del confronto contemporaneo di forma d'onda e spettrogramma: ogni segmento è stato marcato e trascritto con l'alfabeto SAMPA come rappresentato nella figura 10:

[26] Questa omogeneità è resa possibile dal tipo di materiale sonoro utilizzato. Crediamo peraltro che la presenza di rafforzamento sintattico nel sintagma *a destra* sia ininfluente rispetto al ragionamento qui delineato, riservandoci comunque di compiere ulteriori verifiche nel prosieguo della ricerca.

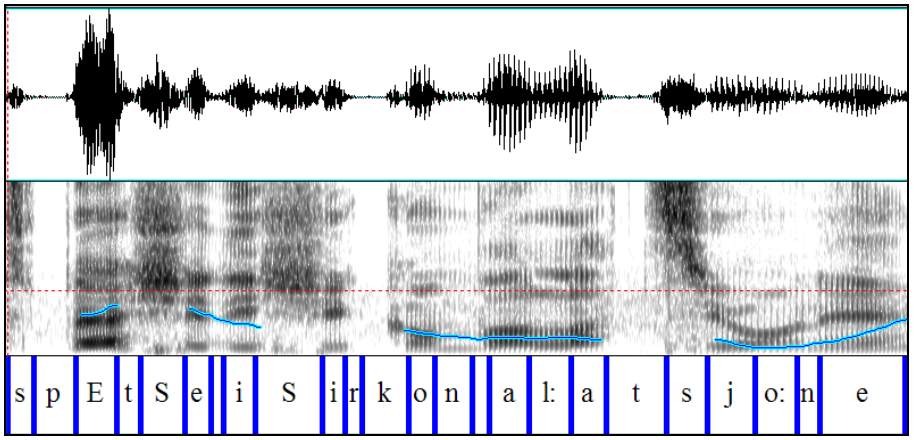


Figura 10: Esempio di etichettatura in Praat: in alto viene rappresentata la forma d'onda, sotto si trovano lo spettrogramma e la sua etichettatura trascritta con l'alfabeto SAMPA

Grazie allo *script* è stato possibile visualizzare gli andamenti formantici dal 5% fino al 95% della durata totale della vocale. Sono stati scelti otto punti di rilevamento anziché dieci, in quanto sia il primo (al 5%) sia l'ultimo (al 95%) sono stati scartati perché maggiormente soggetti ad effetti coarticolatori di natura sia perseverativa che anticipatoria. L'escursione dei valori formantici è stata calcolata sommando i valori assoluti delle differenze, dal 15% al 25% della durata, dal 25% al 35%, e così via, fino all'85%.

Nella tabella 18 sono rappresentati i valori di durata della vocale bandiera /ɛ/ e le relative escursioni formantiche (Δ15-85%) per le tre città oggetto di indagine:

Firenze		**Pisa**		**Livorno**	
Durata (ms)	*Δ15-85% (Hz)*	*Durata (ms)*	*Δ15-85% (Hz)*	*Durata (ms)*	*Δ15-85% (Hz)*
82	84	198	**291**	180	265
78	43	176	170	175	323
104	75	208	346	190	**451**
125	**101**	140	234	131	294
144	80	183	164	143	164

Tabella 18: Durata e Δ15-85% (Hz) della vocale /ɛ/ per le tre località indagate

I dati numerici evidenziano subito quanto le vocali occidentali siano più instabili di quelle fiorentine: le vocali livornesi e pisane presentano valori di durata maggiori rispetto a quelle fiorentine, come emerge dal confronto tra il valore massimo di durata di Firenze, pari a 144 ms, in opposizione ai 190 ms di Livorno e ai 208 ms raggiunti da Pisa. Le vocali nord-occidentali si distaccano da quelle fiorentine anche per quanto concerne l'escursione interna al tracciato formantico che risulta maggiore per i soggetti livornesi ove raggiunge un valore di 451 Hz contro i 291 Hz pisani e i soli 101 Hz fiorentini. Le vocali di Pisa e

Livorno appaiono soggette a un consistente movimento interno, come mostrano gli spettrogrammi nelle Figure 11-13:

(11) Firenze **(12) Pisa** **(13) Livorno**

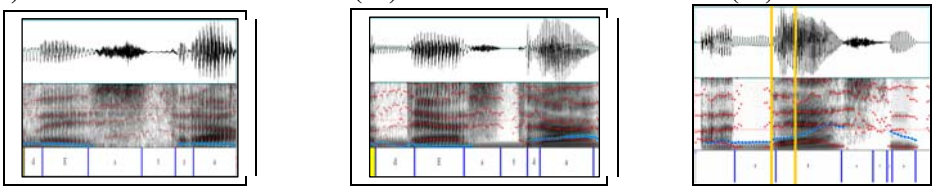


Figure 11-13: Forma d'onda e spettrogramma della parola destra per le città di Firenze, Pisa e Livorno

La vocale medio bassa [ɛ] realizzata dal parlante livornese MP (M) presenta una forma d'onda più disomogenea e meno stabile rispetto agli altri due soggetti; appare diversa anche l'intensità della vocale, maggiore nella sua fase iniziale e minore in quella finale (cfr. Fig.13), quasi come se si trattasse di due diversi segmenti vocalici (per una situazione analoga sugli 'sdoppiamenti vocalici' si veda in proposito il contributo di Marotta, 2003 sul romanesco).

Il maggiore movimento delle vocali livornesi è evidenziato anche dal confronto dei grafici riportati nelle Figure 14-16: sugli assi delle ascisse sono rappresentati gli otto punti di rilevamento delle vocali (dal 15% al 85% della loro durata globale), mentre sugli assi delle ordinate si trova indicato il valore in Hz per F1 e F2.

(14) Firenze **(15) Pisa** **(16) Livorno**

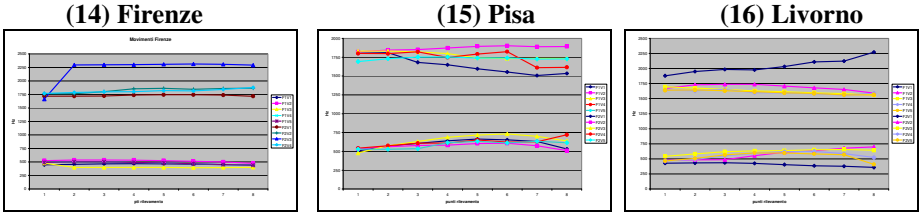


Figure 14-16: Movimenti formantici delle cinque ricorrenze della parola destra per Firenze, Pisa e Livorno

L'andamento delle prime due formanti per i soggetti fiorentini appare molto più costante e lineare (ad eccezione di un valore di F2 per la prima vocale) rispetto a quello livornese che è rappresentato in tutta la sua dinamicità. È proprio la parte centrale (in corrispondenza del quinto e sesto punto – al 55% e al 65% della durata vocalica) che appare caratterizzata da una maggiore instabilità, contravvenendo peraltro a quanto solitamente atteso in questo punto della vocale, di fatto identificato come porzione stabile (*steady state*) della vocale stessa. Anche questo tipo di grafico conferma una differenza interna nell'area nord-occidentale (di cui si è ampiamente discusso al §2): le vocali pisane presentano valori di F1 e F2 maggiori rispetto a quelle livornesi; in particolare, F1 si colloca in una fascia compresa tra i 500 e i 750 Hz e sembra essere la formante maggiormente esposta a variazioni interne che si verificano, come nel caso delle vocali costiere, sempre in corrispondenza del quinto e sesto punto di misurazione. Le vocali pisane sembrano quindi essere più

basse di quelle livornesi,[27] mentre quelle livornesi risulterebbero caratterizzate da due gesti articolatori.[28]

6. CONCLUSIONI

Una ricerca sui fenomeni di riduzione vocalica del parlato spontaneo chiama in causa questioni teoriche e procedurali non facilmente risolvibili, oltre a – non sarà banale ricordarlo – mere difficoltà di misurazione e di rilevamento di confini.

La prima questione riguarda la legittimità del confronto fra parlanti diversi, ovvero la possibilità di considerare differenze sociofonetiche (nella fattispecie geografiche) quelle che potrebbero essere semplici differenze individuali. Il passo successivo della nostra ricerca riguarderà pertanto una adeguata normalizzazione dei dati. Vero è che la dettagliata conoscenza dell'area su stili più controllati rende le nostre osservazioni relativamente meno aleatorie e peregrine, tuttavia gli elevati valori delle deviazioni standard e l'impossibilità oggettiva di controllare il contesto consonantico ci spingono a mantenere un atteggiamento cauto per quanto riguarda gli aspetti più squisitamente dialettologici dell'indagine e – su un piano più generale – a ribadire quanto nelle ricerche di sociofonetica debbano essere tenuti costantemente presenti gli spinosi temi di fonetica generale ancora oggetto di dibattito e di studio.

Un'altra questione solo in parte affrontata riguarda il legame nel parlato spontaneo tra riduzioni timbriche e fattori prosodici. La nostra etichettatura prosodica non è stata – lo ripetiamo – particolarmente sofisticata e necessita senz'altro di miglioramenti. Una più profonda analisi prosodica diventa peraltro particolarmente urgente per l'area occidentale della Toscana ove da più parti è stato rilevato un complesso rapporto tra formanti, durata, andamenti di frequenza fondamentale (Calamai, 2004a; Marotta & Sardelli, 2003; Marotta *et al.*,2004; Gili Fivela, 2008).

La durata risulta essere il parametro più suscettibile a cambiamenti accentuali: sotto accento di frase le vocali sono significativamente più lunghe, sia nel *corpus* totale, sia – con qualche differenza – nelle tre località. Nel *corpus* complessivo, le differenze timbriche riguardano la vocale bassa e le vocali e le vocali medie, che sotto accento di frase diventano più aperte e in generale più periferiche. È comune alle tre località una maggiore sensibilità da parte della prima formante alle variazioni accentuali rispetto alla seconda. Tuttavia, non tutte le vocali sono sensibili allo stesso modo: le vocali bandiera della Toscana occidentale – le medio-basse [ɛ] e, in misura minore, [ɔ] – mostrano di essere sensibili alla variazione accentuale proprio a Pisa e a Livorno, ma non a Firenze, dove restano sostanzialmente stabili. Infine, anche se tutte e tre le località presentano una elevata sensibilità del parametro durata ai cambiamenti relativi all'accento, si registra una diversa gradazione di questa sensibilità: laddove il confronto è significativo, nelle vocali fiorentine l'incremento ha come tetto massimo il 32% (per [a]), mentre nelle vocali pisane raggiunge (peraltro proprio per [ɛ]) il 49%.

[27] In accordo con quanto rilevato su un materiale sonoro completamente differente da Calamai (2004a).

[28] Questa peculiarità livornese, che sembra ricorrere soprattutto nel parlato meno controllato, pare essere condizionata da variabili pragmatiche (quali ad esempio l'enfasi) che andranno meglio indagate.

Anche per la variabile 'posizione della parola' nella frase la durata è il parametro più influenzabile, sia per il sistema tonico che per quello atono (anche se per quest'ultimo gli effetti statistici sono di gran lunga più deboli, anche in considerazione dell'espunzione dal *corpus* delle vocali atone finali, che avrebbero presentato senza dubbio consistenti allungamenti). Il parziale confronto statistico che ha riguardato gli effetti del 'tipo di parola' ha mostrato – come del resto era prevedibile – un maggiore effetto sul vocalismo tonico (vocali più lunghe in parole lessicalmente piene, e tendenzialmente più periferiche).

Il confronto fra sistema atono e sistema tonico mostra, nel *corpus* complessivo, differenze consistenti sul dominio temporale (le vocali toniche sono sempre significativamente più lunghe rispetto a quelle atone) e sulla seconda formante (le vocali toniche sono sempre più periferiche delle corrispondenti atone). Solo nel caso della vocale bassa si raggiunge una elevata significatività statistica per la prima formante, che ha valori molto più elevati sotto accento. Per le altre vocali si registra solo una tendenza ad assumere posizioni più alte e quindi meno aperte sotto accento. Nel parlato semispontaneo il vocalismo atono non è più variabile del vocalismo tonico: le deviazioni standard sono in entrambe i raggruppamenti relativamente elevate. Anche nel *corpus* suddiviso per località le maggiori differenze sono rintracciabili sul piano temporale. Una differenza legata alla variabile 'spazio' è rintracciabile nella consistenza dei cambiamenti, massimi a Firenze, intermedi a Pisa, limitati a Livorno, il cui sistema atono sembra essere nel complesso molto simile a quello tonico.

L'analisi qualitativa dei tracciati formantici condotta sulle ricorrenze della parola *destra* evidenzia come nel *corpus* livornese ci siano maggiori escursioni frequenziali rispetto a quello pisano e fiorentino.

Molto deve essere ancora fatto per capire a fondo il quadro sociofonetico della Toscana occidentale. Crediamo a questo proposito che l'intersecarsi di una prospettiva fonetica con una più strettamente dialettologica si riveli particolarmente utile proprio nell'analisi dei segmenti vocalici: ricerche quantitative offrono una base d'appoggio per osservazioni che rischierebbero di apparire impressionistiche e non verificabili, permettono di rivedere certe tradizionali tassonomie presenti nella letteratura dialettologica (possiamo ancora mantenere l'etichetta di pisano-livornese?), consentono di distinguere opportunamente tra fenomeni fonetici generali e fenomeni che mostrano invece un condizionamento del fattore 'luogo'.

RINGRAZIAMENTI

Desideriamo ringraziare Pier Marco Bertinetto per averci concesso di usare la cabina silente del Laboratorio di Linguistica della Scuola Normale Superiore; a Beat Siebenhaar va la nostra gratitudine per averci concesso di utilizzare lo script di *Praat* da lui creato per la misurazione delle vocali.

Ringraziamo infine i tre revisori anonimi che con la loro attenta e puntuale lettura ci hanno permesso di meglio precisare alcuni punti e di rivedere criticamente alcune nostre posizioni.

7. BIBLIOGRAFIA

Abete, G & Simpson, A. P. (in questo volume), Confini prosodici e variazione segmentale. Analisi acustica dell'alternanza monottongo/dittongo, in *Misura dei parametri. Aspetti tecnologici ed implicazioni nei modelli linguistici*, Atti del 1° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Padova, 2-4 dicembre 2004 (P. Cosi, editor), Torriana (RN): EDK Editore.

Albano Leoni, F., Cutugno, F. & Savy, R. (1995), The vowel system of Italian connected speech, in *Proceedings of XIIIth International Congress of Phonetic Sciences*, Stockholm, August 13-19, 1995 (K. Elenius & P. Branderud, editors), 396-399.

Berruto, G. (1995), *Fondamenti di sociolinguistica*, Roma-Bari: Laterza.

Boersma, P. & Weenink, D. (2005), *PRAAT: doing phonetics by computer*, www.fon.hum.uva.nl/praat/.

Brown, G., Anderson, A., Yule, G. & Shillcock, R. (1984), *Teaching talk*, Cambridge: Cambridge University Press.

Calamai, S. (2001) [2004], Stili a confronto nel parlato toscano (Pisa e Firenze), *L'Italia Dialettale*, 62, 95-125.

Calamai, S. (2004a), *Il vocalismo tonico pisano e livornese. Aspetti storici, percettivi, acustici*, Alessandria: Edizioni dell'Orso.

Calamai, S. (2004b), Vocali fiorentine e vocali pisane a confronto, in *Il parlato Italiano*, Napoli 13-15 Febbraio 2003 (F. Albano Leoni, F. Cutugno, M. Pettorino, R. Savy, editors), Napoli: D'Auria, CD-rom.

Calamai, S. (2007), Per una dialettologia sperimentale, *Rivista Italiana di Linguistica e Dialettologia*, 9, 89-114.

Cerrato, L. & Cutugno F. (1994), Il problema della rappresentazione tempo/frequenza dei fenomeni vocalici dinamici, in *Le vocali: dati sperimentali, problemi linguistici, applicazioni tecnologiche*, Atti delle III^{e} Giornate di Studio del GFS (AIA), Padova, 19-20 novembre 1992 (F. E. Ferrero, & E. Magno Caldognetto, editors), 61-71.

Clemente, G. (2005), La riduzione strutturale e la variazione diafasica nel vocalismo dell'italiano di Palermo, in *Misura dei parametri – aspetti tecnologici e implicazioni nei modelli linguistici*, Atti del 1° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Padova, 2-4 dicembre 2004 (P. Cosi, editor), Torriana (RN): EDK Editore.

Di Benedetto, M.G. (1989), Vowel representation: some observations on temporal and spectral properties of the first formant frequency, *JASA*, 86, 55-66.

Farnetani, E. & Busà, M.G. (1999), Quantifying the range of vowel reduction, *Proceedings of the 14th International Congress of Phonetic Sciences '99*, San Francisco, August 1-7, 1999, 491-494.

Ferrero, F. E. (1972), Caratteristiche acustiche dei fonemi vocalici italiani, *Parole e metodi*, 3, 9-31.

Fourakis, M. (1991), Tempo, stress, and vowel reduction in American English, *JASA*, 90: 1816-1827.

Gili Fivela, B. (2008), Intonation in production and perception: the case of Pisa Italian, in *Memorie del Laboratorio di Linguistica della Scuola Normale Superiore di Pisa*, Alessandria: Edizioni dell'Orso.

Hillenbrand, J., Getty, L.A., Clark, M.J. & Wheeler K. (1995), Acoustic characteristics of American English vowels, *JASA*, 97, 3099-3111.

Lindblom, B. (1963), Spectrographic study of vowel reduction, *JASA*, 35, 1773-1781.

Lisker, L. (1984), On reconciling monophthongal vowel percepts and continuously varying F patterns, *Status Report on Speech Research*, Haskins Laboratories, 79/80, 167-174.

Lo Prejato, M., Clemente, G. & Savy, R. (2004), Su alcuni aspetti della riduzione vocalica nella varietà napoletana, in *Costituzione, gestione e restauro di corpora vocali*, Atti delle XIV^{e} Giornate di Studio del GFS, Università della Tuscia (Viterbo), 4-6 dicembre 2003 (A. De Dominicis, L. Mori, & M. Stefani, editors), Roma: Esagrafica: 183-188.

Giannelli, L. (2000) [1976], *Toscana*, Pisa: Pacini.

Marotta**,** G. (2003), Una nota sulla *lex Porena* in romanesco, *L'Italia dialettale*, 63-64, 87-93.

Marotta, G. & Sardelli, E. (2003), Sulla prosodia della domanda con soggetto postverbale in due varietà di italiano toscano (pisano e senese), in *Voce, canto, parlato. Studi in onore di F. Ferrero*, (P. Cosi, E. Magno Caldognetto & A. Zamboni, editors), Padova: Unipress: 205-212.

Marotta, G., Calamai, S. & Sardelli, E. (2004), Non di sola lunghezza. La modulazione di f0 come indice sociofonetico, in *Costituzione, gestione e restauro di corpora vocali*, Atti delle XIV^{e} Giornate di Studio del GFS, Università della Tuscia (Viterbo), 4-6 dicembre 2003 (A. De Dominicis, L. Mori, & M. Stefani, editors), Roma: Esagrafica: 215-220.

Mioni A. M. (1993), Fonetica e fonologia, in *Introduzione all'italiano contemporaneo. Le strutture* (A. Sobrero, editor), Roma: Laterza, 101-139.

Moon, S. J. & Lindblom, B. (1994), Interaction between duration, context, and speaking style in English stressed vowels, *JASA*, 96, 40-55.

Mooshammer, C. & Geng, C. (2008), Acoustic and articulatory manifestations of vowel reduction in German, *Journal of International Phonetic Association*, 38/2, 117-136.

Padgett, J & Tabain, M. (2005), Adaptive dispersion theory and phonological vowel reduction in Russian, *Phonetica*, 62, 14-4.

Pätzold, M. & Simpson, A.P. (1997), Acoustic analysis of German vowels in the Kiel corpus of read speech, *Arbeitsberichte des Instituts für Phonetik und digitale Spach-verarbeitung Universität Kiel*, 32, 215-247.

Picheny, M.A., Durlach N.I. & Braida L.D. (1986), Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech, *Journal of Speech and Hearing Research*, 29, 434-446.

Rosner, B.S. & Pickering, J. B. (1994), *Vowel perception and production*, New York: Oxford University Press.

Savy R. & Cutugno, F. (1997), Ipoarticolazione, riduzione vocalica, centralizzazione: come interagiscono nella variazione diafasica?, in *Fonetica e fonologia degli stili dell'italiano parlato*, Atti delle VII^{e} Giornate di Studio del GFS, Napoli 14-15 novembre 1996 (F. Cutugno, editor), 177-194.

Savy, R., Clemente, G. & Lo Prejato, M. (2005), Per una caratterizzazione e una misura della riduzione vocalica in italiano*,* in *Misura dei parametri – aspetti tecnologici e implicazioni nei modelli linguistici*, Atti del 1° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Padova, 2-4 dicembre 2004 (P. Cosi, editor), Torriana (RN): EDK Editore, 135-160.

Siebenhaar, B. (*online*): *Online-Einführung in Praat*. Disponibile alla pagina: http://www.germanistik.unibe.ch/siebenhaar/subfolder/PraatEinfuehrung/index.html

Schirru, C. (1994), Aspetti vocalico-temporali dell'italiano in Sardegna. Primi dati sperimentali, Atti delle 4^{e} Giornate di Studio del Gruppo di Fonetica Sperimentale (A.I.A.), Torino, 11-12 novembre 1993 (P. Salza, editor), XXI: 131-140.

Schirru, C. (2000), Peculiarità temporali nel vocalismo dell'italiano in Piemonte: versione integrale, *Italia Dialettale*, 60, 7-24.

Schirru, C. (2003), Caratteristiche vocalico-formantiche dell'italiano in Piemonte, *Bollettino dell'Atlante Linguistico Italiano*, 26/3, 27- 55.

van Bergem, D. (1993), A model of coarticulatory effects on the schwa, *Speech Communication*, 14/ 2, 143-162.

van Bergem, D. (1995), *Acoustic and lexical vowel reduction*, IFOTT: Amsterdam.

van Son, R.J.J.H. (1993), Vowel perception: a closer look at the literature, in *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam,* 17, 33-64.

van Son, R.J.J.H. & Pols, L.C.W. (1990), Formant frequencies of Dutch vowels in a text, read at normal and fast rate, *JASA*, 88, 1683-1693.

van Son R. J. J. H. & Pols, L.C.W. (1992), Formant movements of Dutch vowels in a text, read at normal and fast rate, *JASA*, 92, 121-127.

Vayra, M. (1991), Appunti su un fenomeno di 'centralizzazione' nel vocalismo dell'italiano standard, in *Tra Rinascimento e strutture attuali*, (N. Maraschio, L. Giannelli & T. Poggi Salani, editors), Torino: Rosenberg & Sellier, 195-212.

DIAGNOSTICA FONOLOGICA E DIAGNOSI FONETICA. OSSITONI LUNGHI IN SILLABA LIBERA NEI DIALETTI DI SAMBUCA PISTOIESE (PT)

Lorenzo Filipponio, Nadia Nocchi [1]
Università di Zurigo
filippon@rom.uzh.ch; nocchi_nadia@yahoo.com

1. SOMMARIO

Le isoglosse che compongono la linea La Spezia-Rimini (o, meglio, Carrara-Fano) risultano nell'area appenninica tra Pistoia e Bologna pressoché sovrapposte tra loro: ciò non impedisce però di rilevare anche in questo territorio una scansione dei fenomeni la cui distribuzione può rendere conto della loro cronologia (§2). Un luogo di indagine privilegiato è il territorio di Sambuca Pistoiese (PT), politicamente toscano ma idrograficamente adriatico, costituito da piccole frazioni oramai disabitate, dislocate lungo valli parallele e isolate tra loro. Questa situazione orografica ha di fatto permesso, pur nello spazio ridotto di 77 chilometri quadrati, il mantenimento di caratteristiche linguistiche peculiari e differenziate, in un contesto generale di forte conservatività. Qui sopravvivono (o agonizzano) colonie garfagnine (Treppio), varietà pienamente toscane (Torri), varietà analoghe a quelle dell'alto Appennino bolognese (Pàvana) e altre più marcatamente di transizione (Lagacci, Stabiazzoni, Cavanna, Castello di Sambuca), in cui a un fondo gallo-italico si sono sovrapposti tratti toscani, come è dimostrato, per esempio, dalla sovraestensione della gorgia ai contesti di degeminazione protonica e di mancato raddoppiamento fonosintattico. Abbiamo dunque concentrato la nostra attenzione su queste ultime varietà, per verificare se anche in esse fosse insorta la quantità vocalica distintiva, già operante nelle limitrofe località dell'area bolognese così come a Pàvana.

Vista la resistenza della geminazione postonica, la verifica è stata condotta misurando la durata della vocale tonica degli ossitoni in sillaba libera, priva, come insegna André Martinet (1975), dei condizionamenti dovuti al nesso con la consonante successiva. Molti dialetti gallo-italici hanno infatti sviluppato degli ossitoni secondari con vocale tonica lunga, che formano coppie minime con ossitoni protoromanzi con vocale tonica parametricamente breve. Questi ossitoni secondari si ritrovano, con significative differenze di distribuzione, anche nelle varietà da noi analizzate.

La nostra indagine (§3) ha evidenziato per quelle geograficamente più vicine al toscano, Lagacci e Stabiazzoni, un reintegro degli ossitoni nei parametri dell'italiano standard, che escludono la presenza di vocale tonica lunga finale di parola. Diversa invece è la situazione prospettata dalle parlate di Cavanna e di Castello: in questo caso, infatti, abbiamo registrato la presenza di ossitoni con vocale tonica lunga, costante in contesto di isolamento e finale di frase, saltuaria in contesto interno di frase. Ciò lascia supporre che la lunghezza della vocale tonica degli ossitoni secondari sia foneticamente ammessa, ma non sia integrata strutturalmente, vigendo ancora la quantità consonantica distintiva. Nei casi di presenza in contesto interno di frase di ossitoni con vocale tonica lunga si osserva addirittura la messa

[1] Questo lavoro è frutto di un continuo scambio di vedute tra i due autori. Ciononostante, a fini accademici devono essere attribuiti a Lorenzo Filipponio i §§2.1.-2.4. e 4 e a Nadia Nocchi i §§3.1.-3.3.

in atto di strategie di riparazione, come l'allungamento della consonante iniziale della parola successiva, che oscura la lunghezza vocalica producendo una sorta di raddoppiamento fonosintattico del tutto inaspettato. I dati confermano dunque (§3) un quadro tipicamente di transizione, che aggiunge un nuovo dettaglio alla casistica di Martinet, mostrando un indebolimento in ossitonia del parametro di quantità, che prelude ai successivi sviluppi gallo-italici visibili appena più a nord.

2. DIAGNOSTICA FONOLOGICA: IL QUADRO DIALETTOLOGICO

2.1. Fonologia minima gallo-italica

Il tipo linguistico gallo-italico, che qui consideriamo seguendo Pellegrini (1992) come italo-romanzo settentrionale escluso il veneto, è individuabile dal punto di vista fonetico-fonologico da una serie di caratteristiche comuni scaturite dalla progressiva trasformazione della struttura di parola latina.

Per quanto concerne il vocalismo tonico, bisogna innanzitutto tenere conto dell'allungamento di vocale tonica in sillaba libera, fenomeno protoromanzo (e quindi esteso ben oltre i limiti del gallo-italico: Haudricourt & Juilland, 1949: 32ss.; Lüdtke, 1956: 134ss.; Weinrich, 1958: 181) che ha limitato le strutture sillabiche toniche ereditate dal latino alle varianti $ˈ(C)Vː$ ~ $ˈ(C)VC$. Da questo processo sono stati parametricamente esclusi, almeno in area gallo-italica così come in toscano,[2] gli ossitoni e i monosillabi terminanti per vocale presenti *ab origine* nelle lingue romanze (Rohlfs, 1966: § 9). Nei dialetti gallo-italici si è in seguito manifestato un fenomeno di compensazione ritmica nei proparossitoni, vale a dire una riduzione della quantità della vocale tonica indotta dalla presenza di cospicuo materiale sillabico atono a destra della sillaba accentata (Marotta, 1985: 27ss.; Loporcaro, 2005: 104).[3] A esemplificare questa sequenza, si osservino i casi di *PĔTRA e *TĔPIDU. Nel primo caso, l'esito odierno bolognese (con metatesi) è [ˈpreːda], dove [eː] è il risultato regolare dell'allungamento di Ĕ in sillaba aperta; nel secondo caso, l'esito odierno bolognese è [ˈtavd], che presuppone una trafila, confermata dai dati delle varietà appenniniche corrispondenti, < [ˈtɛvd] (medio Appennino bolognese) < [ˈtevd] (Porretta Terme), il cui profilo timbrico non può che presupporre un *[ˈteːvido], visto che un trattamento originario di sillaba chiusa (che postulerebbe nei proparossitoni la compensazione ritmica precedente all'allungamento di vocale tonica in sillaba libera) avrebbe dato come esito un *[ˈtɛvːido] (con presumibile geminazione anetimologica) per arrivare a un odierno *[ˈtɛːvd] (con allungamento secondario: cfr. Uguzzoni, 1975: 69ss.), esattamente come il regolarissimo *PĔCTINE > [ˈpɛtːne] (alto Appennino bolognese) > [ˈpɛːten] (Bologna), che del trattamento di sillaba chiusa manifesta appunto i segni. Tale trattamento è stato invece riservato agli ossitoni e monosillabi primari di cui abbiamo detto sopra (cfr. Loporcaro, 1997: 71), come mostra l'esito bolognese odierno di RĒX, [ra] (Coco 1970, §20), con Ē intercettata nel processo di abbassamento timbrico delle vocali toniche medioalte brevi già visto in *TĔPIDU.

[2] In Loporcaro (1997: 70-72) viene evidenziato il fatto che alcune varietà meridionali, così come – fuori dall'Italia – alcune varietà gallo-romanze settentrionali, non mostrano tale esclusione parametrica.

[3] La compensazione ritmica ha innescato successivamente un fenomeno più generalizzato di riduzione della quantità vocalica che si è andato estendendo in alcune varietà (come il milanese) ai parossitoni e in altre (come il bergamasco) anche agli ossitoni secondari.

Per quanto concerne il vocalismo atono, si osserva un fenomeno generalizzato di progressiva scomparsa, che riguarda in una prima fase le vocali atone interne (sincope) e successivamente quelle finali (apocope). Sia sincope sia apocope si sono manifestate per gradi: specialmente nel caso dell'apocope, la situazione odierna dei dialetti italo-romanzi settentrionali presenta ancora una casistica estremamente articolata, che lascia trasparire i diversi stadi di avanzamento del fenomeno (cfr. Loporcaro, 2005-6). Analoga, ma meno trasparente alla luce dei dati odierni, è la situazione della sincope.[4]

Per quanto concerne infine il consonantismo, si ha a che fare con i noti fenomeni di lenizione e degeminazione, la cui cronologia relativa è stata ampiamente discussa – cfr. a titolo esemplificativo Martinet (1955: 280) e Weinrich (1958: 145-146) – e sui quali in questa sede non ci soffermiamo oltre.

2.2. La posizione di Sambuca Pistoiese nel quadro gallo-italico

.

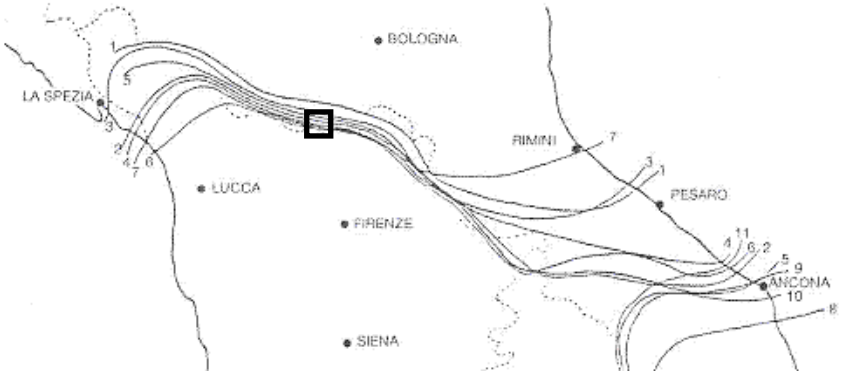
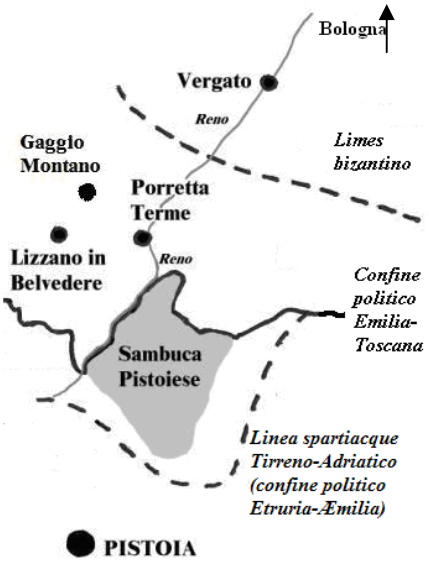


Figura 1: Sambuca Pistoiese rispetto alla linea La Spezia-Rimini (o Carrara-Fano)

La succitata serie di fenomeni ha dunque contribuito in larga a parte a conformare l'assetto fonologico delle varietà gallo-italiche. In un simile quadro, l'osservazione delle cosiddette aree marginali o laterali può contribuire a ricostruire nel dettaglio le tappe di questa evoluzione: al riguardo, il territorio di Sambuca Pistoiese, in provincia di Pistoia, è estremamente interessante. Siamo pressoché a cavallo del fascio di isoglosse che separa i

[4] Gli stadi di avanzamento della sincope dipendono principalmente dalla tipologia delle consonanti precedenti e seguenti la vocale passibile di caduta: i contesti più favorevoli al verificarsi del fenomeno, che difatti sono anche i primi a esserne interessati, sono quelli con sonorante prevocalica e ostruente occlusiva postvocalica (cfr. casi già latini del tipo CALĬDUS > CALDUS); a posizioni invertite, il contesto è sfavorevole, e viene interessato dalla sincope solo successivamente (cfr. per il gallo-romanzo la dettagliata cronologia di Richter, 1934): come dimostrato in Filipponio (in corso di stampa[a]), questa fenomenologia è perfettamente spiegabile facendo ricorso alle leggi di preferenza sillabica di Vennemann (1988). Una ricognizione sul trattamento della sincope nei proparossitoni etimologici nell'area dell'Appennino bolognese è reperibile in Filipponio (2007[a]: Tab. 3: 96, si leggano [Je], [E] come [je], [e]; nota 11: 95, si legga «precedente» come «successiva»).

dialetti gallo-italici dal toscano (cfr. Fig. 1), già sul versante settentrionale dello spartiacque appenninico ma in un'area che dall'Alto Medioevo in poi è rimasta sotto il controllo di Pistoia.[5] Il territorio è articolato in tre valli principali, separate da alture piuttosto impervie, in cui scorrono il fiume Reno e due suoi affluenti, i torrenti Limentra Occidentale e Limentra Orientale.[6]



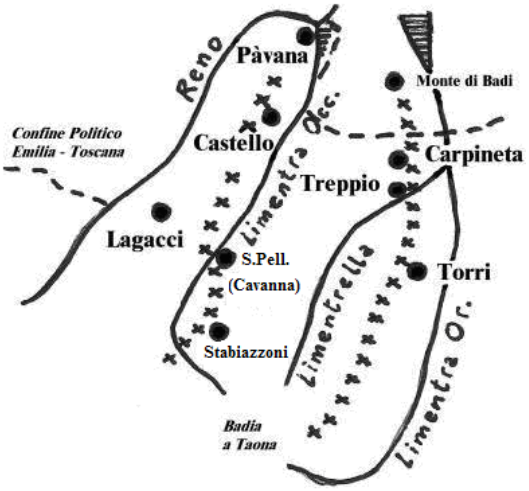


Figure 2 e 3: Il territorio di Sambuca Pistoiese[7]

Muovendosi tra queste valli, già a una prima ricognizione impressionistica un orecchio appena allenato avverte una situazione linguisticamente ibrida, che può essere interpretata come una sovrapposizione toscana a un fondo gallo-italico: ciò è verificabile osservando, per esempio, il fatto che a Castello si rilevi una sovraestensione della gorgia (= sovrapposizione toscana) in contesti in cui o manca il raddoppiamento fonosintattico, come, per esempio, [da ˈxwesto], 'da questo', o è intervenuta una degeminazione protonica (= fondo gallo-italico: cfr. (3) di Tab. 1), come [d aˈxɔɾdo], 'd'accordo' {audio 1}. Bisogna però

[5] A partire per l'esattezza dall'avanzata longobarda successiva all'occupazione di Firenze (593 d.C.), che ricacciò i Bizantini al di là di una linea difensiva situata in pieno medio Appennino bolognese (il *limes* di Fig. 2), a sud della quale il controllo pistoiese si mantenne fino al 1219 (Lodo di Viterbo), quando vennero stabiliti i confini che ancora oggi dividono la provincia di Bologna da quella di Pistoia. Tali confini hanno lasciato alla città toscana l'area di Sambuca, a nord di quello spartiacque appenninico che in epoca imperiale determinava il confine tra la *Regio VII Etruria* e la *Regio VIII Æmilia*.

[6] Una quarta valle, pressoché disabitata, è quella del Limentrella, affluente del Limentra Orientale (cfr. Fig. 3).

[7] I due tracciati a crocette della Figura 3 indicano le principali vie di comunicazione transappenniniche medievali sulla direttrice Bologna-Pistoia. Tutte le località menzionate, tranne Monte di Badi (BO), sono frazioni del comune di Sambuca Pistoiese.

tenere conto del fatto che l'articolazione dialettale di questo territorio è estremamente frammentata. Il caso più eclatante, e più discusso, è sicuramente quello di Treppio (e della limitrofa località di Carpineta, cfr. Fig. 3), il cui dialetto presenta caratteristiche che hanno fatto di volta in volta pensare al relitto preindoeuropeo o, più verosimilmente, alla colonia garfagnina:[8] quindi, ai fini della presente ricerca, tesa a evidenziare alcuni momenti della formazione del tipo fonologico gallo-italico alla luce delle aree laterali, il treppiese manifesta una problematicità che ci ha indotto a escluderlo dal novero delle varietà da analizzare. Abbiamo dunque ristretto la nostra attenzione: a una località situata sul pendio che scende verso la valle del Reno (attraversata dal 1864 dalla linea ferroviaria Bologna-Pistoia), Lagacci; a quattro località situate nella valle del Limentra Occidentale (che è il canale principale di comunicazione, oggi attraversato dalla S.S. Porrettana), da nord verso sud Pàvana, la già citata Castello di Sambuca, San Pellegrino (fraz. Cavanna) e Stabiazzoni; a una località in prossimità del crinale che divide la valle del Limentrella da quella del Limentra Orientale, ma che insiste su quest'ultima per quanto concerne le comunicazioni (cfr. Fig. 3), ovvero Torri.

Consideriamo ora come i fenomeni elencati al §2.1. si manifestano nelle varietà delle frazioni sambucane; ne risulterà una sorta di patente fonologica della gallo-italicità aurorale di questi dialetti:

	(+ = presente; **-** = assente)	**Torri**	**Lagacci**	**Castello Cavanna Stabiazzoni**	**Pàvana**
1	ĕ, ŏ in sill. lib. > [eː], [oː]	**-** [ˈfwɔxo]	**+** [ˈfoːgo]	**+** [ˈfoːgo]	**+** [ˈfoːgo]
2	lenizione intervocalica	**-** [aˈʃeːto]	**+/-** [ˈpegora] [aˈʃeːto]	**+** [aˈʒeːdo]	**+** [aˈʒeːdo]
3	degeminazione protonica	**-** [kaˈtːiːva]	**+** [doˈniːna]	**+** [kaˈtiːva]	**+** [gaˈliːna]
4	comp. ritmica nei proparossitoni	**-** [ˈviːɸera]	**+/-** [ˈvipˑera] [ˈtjɛpido][9]	**+** [ˈvipˑera]	**+** [ˈvipˑara]
5	apocope dopo *-n-* postonica	**-** [ˈbwɔno] [ˈfjɛno]	**-o > -e** [ˈboːne] [ˈfeːne]	**+** [ˈbõː], [ˈfẽː]	**+** [ˈbõː], [ˈfẽː]
6	sincope voc. atona interna	**-** [ˈstɔˑmaxo] [ˈtjɛɸido]	**-a-, -i- > -e-** [ˈʃtomˑego] [ˈmanˑego]	**-a-, -i- > -e-** [ˈʃtomˑego] [ˈtevedo]	**+** [ˈʃtomˑgo] [ˈtevdo]

Tabella 1: Alcuni caratteri fonetico-fonologici delle varietà parlate a Sambuca Pistoiese

[8] Lo *shibboleth* treppiese è il passaggio di *l-*, *-ll-* a [ɖ]-, -[ɖ]-, ben noto in Garfagnana. Una piccola ricognizione sul problema, con relative indicazioni bibliografiche, è contenuta in Filipponio (2008a).

[9] L'esito con dittongazione romanza di *TĔPIDU si registra ([ˈtjɛβdo]) anche sulla riva sinistra della valle del Reno, nel territorio già emiliano di Granaglione (BO).

(1) è la peculiare evoluzione timbrica gallo-italica di Ĕ, Ŏ conseguenza dell'allungamento di vocale tonica in sillaba libera; (3) richiama l'attenzione sul fatto che la degeminazione consonantica si è manifestata primariamente in protonia; (5) si riferisce al primo contesto in cui si è verificata apocope della vocale finale atona, con nasalizzazione delle vocale tonica.

Da questa tabella risulta che la frazione di Torri attesta una varietà pienamente toscana, come indicano la presenza della dittongazione romanza, l'assenza della compensazione ritmica, la gorgia toscana, la preservazione del consonantismo e di tutte le vocali atone (cfr. Filipponio 2007c),[10] caratteristiche alle quali va aggiunta la presenza del raddoppiamento fonosintattico, assente nel resto del territorio sambucano. Ciò ci induce a scartare, ai fini della nostra ricerca, questa varietà, per concentrarci sulle valli del Reno e del Limentra Orientale. Non mancano anche in questo caso alcuni dati interessanti: osservando (4), si può intuire che nella parlata di Lagacci numerosi proparossitoni sfuggano alla compensazione ritmica, forse per una maggiore interferenza col toscano (cfr. per i dati completi Filipponio, in corso di stampa b). Per quanto concerne (5), ancora Lagacci mostra un tratto estremamente interessante che può essere aggiunto alla casistica dell'apocope tratteggiata in Loporcaro (2005-6): in un contesto tra i primi a essere interessati dal fenomeno, e cioè dopo *-n-* postonica, la vocale finale *-o* è stata ricostruita come *-e* probabilmente attraverso una fase intermedia *-ə*; stesso destino può essere ipotizzato per l'altra vocale media *-e*, come *-o* primariamente esposta all'indebolimento e alla caduta, con preservazione, forse, del morfema di femminile plurale (cfr. Loporcaro, 2005-6: *passim*). Il punto (6) disvela un'altra caratteristica di buona parte delle varietà sambucane, cioè l'indebolimento, che non arriva a caduta, delle postoniche interne *-i-* e *-a-*, attestate come *-e-*, per le quali non sarà improbabile anche in questo frangente ipotizzare l'esistenza uno stadio intermedio *-ə-*.[11]

2.3. Ossitoni con vocale tonica lunga in sillaba libera

La casistica dei fenomeni descritti nel §2.1. ha portato in numerose varietà gallo-italiche all'insorgenza della quantità vocalica distintiva. Già il dialetto di Lizzano in Belvedere, nell'alto Appennino bolognese, così come quello della vicina Porretta, poco più a nord di Sambuca sulla direttrice della valle del Reno e della S.S. Porrettana, presentano questo assetto. Bisogna dunque verificare se è possibile dire lo stesso, alla luce di quanto visto in Tab. 1, anche per le varietà sambucane, ancora più marginali rispetto a quelle di Lizzano e Porretta: la diagnostica fonologica fornita nella tabella, però, non è sufficiente,

[10] Questo breve scritto ha scatenato in ambito locale una polemica piuttosto accesa; testimonianze storiche, infatti, lasciano intravedere il fatto che Torri sia stata ricolonizzata nel XVI secolo da una comunità proveniente dall'alto Appennino modenese-reggiano, il che ha fatto pensare a una storia dialettologica in tre fasi: sambucana, emiliana, toscana. Se la toponomastica testimonia lo strato sambucano, e le fonti testimoniano la presenza modenese-reggiana, corroborata anche dal fatto che, in sintonia con alcune aree alto-modenesi, il termine torrigiano per 'ragazza' è *guarzetta*, la toscanità linguistica appare a Torri strutturalmente profonda e radicata, e solidale con altre aree montane della provincia nel restituire caratteristiche tipiche di una varietà pistoiese antica (cfr. Filipponio, 2007c; 2008b; Vitali, 2009).

[11] Questo indebolimento non interessa indistintamente tutte le *-i-* e *-a-* postoniche interne, ma dipende dal saldo di forza consonantica delle consonanti che precedono e che seguono la vocale atona (cfr. la nota 4).

trattandosi oltretutto di dialetti in cui la geminazione postonica si manifesta ancora salda.[12] In un simile scenario, dunque, è necessario fare ricorso al test proposto da Martinet (1975: 205) per l'individuazione dell'esistenza di quantità vocalica distintiva, applicato da Loporcaro *et al.* (2006: 512-514) proprio al dialetto di Lizzano. In sostanza, si tratta di verificare la durata della vocale tonica degli ossitoni e dei monosillabi uscenti in vocale non presenti *ab origine* nelle lingue romanze, ma formatisi in seguito a fenomeni di apocope della sillaba finale atona. In diverse varietà gallo-italiche questi ossitoni secondari presentano una vocale tonica lunga, e formano coppie minime con altri ossitoni: in lizzanese, per esempio, il participio passato [kanˈta] 'cantato' si oppone alla seconda persona plurale del presente indicativo [kanˈtaː] '(voi) cantate'. Come suggerisce Martinet, l'assenza di consonanti postoniche che possano opacizzare il contesto rende la verifica di queste alternanze una prova pressoché incontrovertibile della presenza di opposizione di quantità vocalica.[13] Ora, il dialetto di Pàvana, sicuramente il più 'emiliano' nel lotto delle parlate sambucane (cfr. Guccini, 1998), mostra un quadro in piena sintonia con quello di Lizzano (cfr. anche *pé* [pe], 'piede' ~ *pée* [peː], 'piedi'; Guccini, 1998: 72) e dunque può essere pacificamente ascritto al novero delle varietà in cui la geminazione consonantica ha mera consistenza fonetica. Resta dunque da verificare la situazione nel resto della valle del Limentra Occidentale e nella valle del Reno. Per questo abbiamo alfine concentrato le nostre indagini sulle quattro località di Castello, Cavanna, Stabiazzoni e Lagacci.

2.4. La formazione degli ossitoni secondari

Ovviamente, a causare la formazione di ossitoni secondari è la perdita di materiale atono a destra della sillaba accentata. Facendo riferimento a una varietà limitrofa più volte chiamata in causa, e comunque conservativa nel quadro gallo-italico, l'inventario più dettagliato di forme ossitone secondarie è quello di Malagoli (1930) per il lizzanese. Gli ossitoni secondari e i monosillabi con vocale tonica lunga ivi riportati forniscono il materiale di partenza sia per un *corpus* su cui basare l'indagine sul campo, sia per una tabulazione che permetta di verificare la presenza o meno di queste forme nelle varietà qui sotto esame.

Prima dunque di misurare la durata delle vocali toniche in questi ossitoni, è possibile osservarne la distribuzione. Il fatto interessante, e non atteso, è che anche questo quadro, come quello mostrato in Tab. 1, presenta esiti differenziati all'interno del territorio

[12] Sulla quantità vocalica distintiva nel lizzanese cfr. Loporcaro *et al.* (2006), che rilevano la resistenza, ormai non più fonologicamente pertinente, di consonanti geminate postoniche (anche anetimologiche, in caso di compensazione ritmica nei proparossitoni); i dati contenuti in Filipponio (in corso di stampa[a]) presentano un quadro in cui la geminazione consonantica appare molto più erosa. Non è da escludere che questa discrasia derivi da una diversa modalità di elicitazione dei dati: il differente comportamento fonetico a seconda della variazione del contesto può essere visto come un'ulteriore conferma della non pertinenza fonologica di questa geminazione.

[13] Per i dialetti del Frignano (Appennino modenese) Uguzzoni *et al.* (2003) hanno ipotizzato il passaggio da una opposizione di quantità vocalica a una di taglio sillabico, tipica di alcune lingue germaniche; simile evoluzione prevedrebbe però la presenza parametrica di vocali toniche lunghe negli ossitoni in sillaba libera, come è in effetti in inglese o in tedesco (cfr. Martinet, 1966; Vennemann, 2000), ma non nel Frignano, né in bolognese, né tantomeno nei dialetti sambucani.

sambucano. Lagacci, che anche nel prospetto precedente, ad eccezione del caso di Torri, mostrava la maggiore distanza dalla pienezza gallo-italica e le maggiori interferenze col toscano, presenta sulla base del nostro piccolo *corpus* il numero più ridotto di casi di ossitonia secondaria; le tre frazioni della valle del Limentra Occidentale si dispongono regolarmente da sud verso nord avvicinandosi progressivamente al lizzanese: la coerenza territoriale è dunque garantita. Con essa, però, si può garantire anche quella strutturale; sotto questo punto di vista, una riflessione approfondita su questi dati aprirebbe pagine di linguistica storica romanza che ci riserviamo di sfogliare in altra sede: ci si accontenti pertanto di alcune osservazioni cursorie, ricordando che il segno + in Tab. 2 indica la presenza degli ossitoni secondari ma non dice alcunché circa la lunghezza della loro vocale tonica, che analizzeremo nei prossimi paragrafi.

	forma	**cond. partenza**	**LIZZANO**	**CAST**	**CAV**	**STAB**	**LAG**
2c	*fratelli*	-ĔLLI	[fraˈdeː]				
	coltelli	-ĔLLI	[korˈteː]				
2b	*stivali*	*-ĀLI	[ʃtiˈvaː]	+			
	paioli	-ŎLI	[paˈroː]	+			
	badili	*-ĪLI	[baˈdiː]	+			
1b	*mortai*	< *-ari* < -ĀRI[14]	[morˈtaː]	+			
2a	*(voi) cantate*	*-adi* < -ĀTIS	[kanˈtaː]	+	+		
	(voi) vedete	*-edi* < -ĒTIS	[veˈdeː]	+	+		
	(voi) dormite	*-idi* < -ĪTIS	[dorˈmiː]	+	+		
2b	*figli*	-ŎLI	[fjoː]	+	+	+	
2a	*piedi*	*PĔDI	[peː]	+	+	+	
1a	*suoi* (tonico)	*SŎI	[soː]	+	[+]	[+]	
	miei (tonico)	MĔI	[meː]	+	+	+	
	tuoi (tonico)[15]	*TŎI[16]	[toː]	+	+	+	
	(tu) vuoi	< *vuoli*	[voː]	+	+	+	+
	(io) farei	*-èi* (< *ebbi*)	[faˈreː]	+	+	+	+
	lui	*ILLUI	[luː]	+	+	+	+
	lei	*ILLEI	[leː]	+	+	+	+
	sei	SEX	[seː]	+	+	+	+
	due	**dui*[17]	[duː]	+	+	+	+

Tabella 2: Distribuzione degli ossitoni secondari nelle varietà indagate

[14] Il rifacimento analogico sul singolare è subentrato al plurale in *-ari,* regolare in toscano fino al Trecento (Rohlfs, 1966: §284; 1969: §1072).
[15] In proclisi si ha vocale tonica breve ([mɛ], [tɔ], [sɔ]: Malagoli, 1940: §15).
[16] Per *SŎI e *TŎI ci affidiamo qui alla ricostruzione di Rohlfs (1968: §427) basata sulla *vulgata* Monaci-D'Ovidio, contro la quale si schierò *in primis* Castellani (1952: 75ss.). Sull'argomento cfr. la ricognizione di Barbato (in stampa).
[17] Cfr. Malagoli (1930, §48).

Stando ai dati qui presentati, da leggersi dal basso verso l'alto, si possono osservare due tappe principali nel processo di formazione degli ossitoni secondari: (1) lo iato primario di vocale tonica davanti a *-i* e (2) lo iato secondario di vocale tonica davanti a *-i*, attraverso, secondo Malagoli (1930: §48), una condizione intermedia di dittongo discendente poi ridotto: -ˈV$i# > -ˈV̯i# > -ˈV#.[18] Nella varietà di Lagacci, soltanto il contesto (1) ammette la formazione degli ossitoni secondari, anche se i pronomi possessivi singolari restituiscono forme toscane (*mia*, *tua*, *sua*) che interferiscono anche nelle risposte degli informatori di Stabiazzoni e Cavanna.[19] Seguono i contesti di iato secondario, scanditi in tre momenti: (2a) iato secondario dovuto a scomparsa di *-d-* (< *-t-*) intervocalica (Rohlfs, 1966: §201; §216);[20] (2b) iato secondario dovuto a palatalizzazione e successiva scomparsa di *-l-* davanti a *-i* di plurale (Rohlfs, 1966: §221); (2c) iato secondario dovuto a palatalizzazione e successiva scomparsa di *-ll-* davanti a *-i* di plurale (Rohlfs, 1966: §233). In questo caso l'articolazione degli esiti è meno prevedibile, ma tutt'altro che inspiegabile: si guardi alla forma ossitona *fió*,[21] ben attestata in tutta la valle del Limentra Occidentale, che in una scansione perfettamente regolare ci saremmo aspettati attestata solo a Castello in compagnia di *stivà*, *paró*, *badì*. Non escluderemmo qui a priori che la referenza [+umano] della parola in questione, e di conseguenza la possibilità di un suo utilizzo in contesti prosodicamente marcati come quelli vocativi, abbiano favorito il determinarsi di condizioni di ossitonia. Per quanto concerne (2a), sembra che una isoglossa separi Cavanna da Stabiazzoni, il cui informatore ricorda per la seconda persona plurale dell'indicativo presente forme toscane,[22] mentre, a parità di contesto fonetico, è attestato anche a Stabiazzoni il tipo *pé*. Per il resto, l'ossitonia secondaria sotto le condizioni in (2b) è regolare soltanto a Castello,[23] che invece sotto (2c) presenta lo stadio intermedio [fraˈdɛʎːi], [korˈtɛʎːi], in accordo con Pàvana (Guccini, 1998: 54), che in questo frangente si differenzia da Lizzano.

---

[18] Questo fenomeno di riduzione del dittongo discendente può essere messo in parallelo con quello analogo, interno di parola, descritto per il fiorentino antico da Castellani (1952: 106ss.: *preite* > *prete*; *aitare* > *atare*; *voito* > *voto*, ecc.).

[19] Da qui il segno [+] nella Tab. 2. Non è sempre facile in questi casi ascrivere i fenomeni a interferenza recente col toscano o a una situazione consolidata e connaturata alla condizione geolinguistica del dialetto analizzato.

[20] Per le forme della seconda persona plurale dell'indicativo presente cfr. Rohlfs (1968, §531).

[21] Che, come altre parole menzionate nel §2.4., non trascriviamo foneticamente perché non abbiamo ancora specificato la durata della vocale tonica e il suo eventuale peso fonologico.

[22] Anche in questo caso si pone il problema se si tratti di forme originarie oppure dovute a una rimonta toscana di cronologia alta o veicolata in tempi più recenti dalla sovrapposizione dell'italiano (cfr. la nota 19). A Lagacci queste forme suonano *cantaddi*, *vededdi*, *dormiddi*.

[23] La conservazione di *-i* in *mortai* (cfr. la nota 14) deve essere considerata un toscanismo. Secondo Agostiniani (1989: 24), nel parlato di Toscana il mantenimento in iato di *-i* postvocalico morfema di plurale maschile è dovuto al fatto che, al contrario di quello che accade in casi del tipo *ci anda' di corsa*, in cui il passato remoto *andai* viene analizzato come /and+a+i/, qui il segmento costituisce da solo un morfema flessivo (/morta+i/). Per questa parola (in Tab. 2, 1b) soltanto l'informatore di Castello riporta l'esito ossitono analogo al lizzanese. A Bologna si registrano per 'mortaio' *pistån* (Ungarelli, 1901: 214) e

Come abbiamo detto, la presenza di vocale tonica lunga, certa per il lizzanese, non fa (ancora) testo per il resto della tabella. Lo fa invece il timbro della vocale tonica: se postuliamo per le varietà di Sambuca condizioni emiliane di partenza, possiamo allora rifarci a quanto dice Malagoli (1930, §47ss.) sul fatto che la posizione di iato primario davanti a *-i* autorizza il trattamento di sillaba libera.[24]

Dopo questa breve rassegna, non resta che dedicarsi alla ricerca di eventuali vocali toniche lunghe negli ossitoni secondari in sillaba libera.

3. DIAGNOSI FONETICA: L'INCHIESTA SUL CAMPO

3.1. Materiali e metodi

I dati, raccolti dal primo Autore, riguardano, come illustrato nel §2.2, le località di Castello, Cavanna e Stabiazzoni nella valle del Limentra Occidentale e quella di Lagacci, nella valle del Reno.

Sono stati registrati otto informatori, sei maschi e due femmine, la cui dialettofonia viene però esercitata in un territorio in cui la condizione del dialetto è estremamente periclitante, erosa qualitativamente dalla pressione dell'italiano standard (qui veicolato dalla rimonta toscana) e quantitativamente dallo spopolamento tipico, nel secondo dopoguerra, di tutte le aree montane.[25] Le frazioni oggetto della nostra inchiesta contano infatti poche decine di abitanti; a Castello si contano poche unità. In seguito a una ricognizione del materiale raccolto, abbiamo deciso di concentrare l'analisi spettrografica sui dati di quattro parlanti, tutti di sesso maschile, di cui riportiamo iniziali, età e professione: per Castello, SC (84), postino in pensione; per Lagacci, RG (75), ex-operaio presso la cava di Campo Tizzoro nei pressi del Monte Orsigna (a 13km da Lagacci in direzione S. Marcello Pistoiese); per Cavanna, GG (68), artigiano in pensione; per Stabiazzoni, infine, AT (63), artigiano. Tutti risiedono in pianta stabile nelle località di origine.

Gli incontri con gli informatori sono avvenuti principalmente a casa degli stessi o in luoghi di aggregazione del paese. Per ottenere un risultato più fedele possibile, non si è mai cominciato *ex abrupto*, ma si è preferito scegliere un approccio 'emico',[26] concedendo un

---

murtæ̀l (Ungarelli, 1901: 185), con scambio di suffisso (DEI: 2513, s.v. *mortale*[2]) tutt'altro che infrequente sia in Italia settentrionale sia nell'area altomeridionale (cfr. AIS: c. 978Cp). È fondamentale osservare che, pur interessando in parte le stesse parole, il fenomeno di elisione di *-i* postvocalico analizzato da Agostiniani (1989) per il toscano non è lo stesso fenomeno trattato in queste pagine e che conduce alla formazione degli ossitoni secondari, ma è un tratto meramente superficiale. Prova ne è il fatto che nel nostro caso la caduta di *-i* si produce anche in isolamento o in posizione finale di frase, mentre condizione necessaria per il verificarsi del fenomeno nel parlato di Toscana è la presenza di un qualsiasi segmento successivo.

[24] Cosa che rende opachi eventuali effetti di metafonia a contatto sul timbro delle vocali medio-basse, che sono invece trasparenti negli esiti lizzanesi, ma non sambucani (cfr. Tab. 2), del tipo [fraˈdeː], [korˈteː].

[25] Il comune di Sambuca Pistoiese contava 7.167 abitanti nel 1911, 4.668 nel 1951, 1.604 nel 2001 (fonte: ISTAT). Di questi, una percentuale consistente vive a Pàvana, la località più vicina a Porretta e alla ferrovia.

[26] Secondo Carpitelli & Iannaccaro (1995: 99), 'emico' si riferisce al modo in cui si intessono rapporti di conoscenza con membri di una determinata comunità (e quindi

lasso di tempo alla presentazione e alla conversazione: del resto, già Jaberg & Jud (1987 [1928]: 242ss.) sottolineavano come "niente in un'inchiesta è più d'ostacolo della mancanza di libertà interiore dell'intervistat"».

Per assemblare il questionario è stato usato come riferimento il repertorio lizzanese in appendice al saggio fonetico di Malagoli (1930: 188-196; cfr. Tab. 2), da cui sono stati estratti tutti gli ossitoni in sillaba libera con vocale tonica lunga; a questi sono stati aggiunti ossitoni monosillabi e bisillabi con vocale tonica breve (del tipo *qua*, *giù*, *caffè*, *città*),[27] a comporre un *corpus* di 57 parole che sono state fatte pronunciare in isolamento e in due serie di frasi, la prima in cui gli ossitoni compaiono in posizione finale di frase, la seconda in cui gli ossitoni sono stati collocati in posizione interna, evitando i confini di costituenti sintattici maggiori. Per la selezione delle frasi si è seguito lo stesso principio di pertinenza diafasica utilizzato nel caso della scelta delle parole: seppur ragionevolmente brevi, esse hanno cercato infatti di riprodurre situazioni quotidiane o legate alla vita rurale, affinché gli informatori trovassero estremamente naturale ripeterle nel proprio dialetto. Se ne forniscono alcuni esempi nella Tab. 3:

Ossitoni secondari	**Ossitoni primari**
(io) farei	*caffè*
Se avessi voglia, lo farei volentieri	Il nonno beve caffè corretto
(voi) cantate	*città*
Voi cantate tutte le domeniche	Siamo andati in città con le ragazze
due	*virtù*
Avete preso due pentole per fare la marmellata	La calma è la virtù dei forti

Tabella 3: Esempio delle frasi del *corpus*

Le frasi del *corpus* sono state presentate in italiano dall'intervistatore e fatte ripetere per tre volte in dialetto dall'intervistato, in modo da ottenere un numero adeguato di *items* in una forma elicitata di parlato in cui la competenza linguistica del parlante potesse garantire la non artificiosità del fenomeno indagato. Nessun informatore ha mostrato di avere coscienza del tipo di indagine che si stava svolgendo: i più sorvegliati hanno rivelato una certa sensibilità metalinguistica, limitata però agli aspetti lessicali – attraverso il raffronto con altri dialetti – e alla superficie fonetica.

Tutte le interviste sono state effettuate con un registratore digitale Fostex FR2LE ed un microfono a cravatta modello Sennheiser MKH2.[28] L'acquisizione del segnale è avvenuta a una frequenza di campionamento di 44100 Hz a 16 bit di ampiezza; successivamente, nella fase di riversamento per l'analisi su pc, il materiale sonoro analizzato è stato portato ad una frequenza di campionamento pari a 22050 Hz mantenendo lo stesso numero di bit.

---

rappresentanti di una determinata cultura) e, in particolare, a un tipo di approccio che cerchi di tenere conto della loro peculiare visione del mondo.

[27] Per evitare eventuali fraintendimenti, segnaliamo qui che nel testo le definizioni di 'ossitoni (presunti) lunghi' e 'ossitoni secondari', così come quelle di 'ossitoni (presunti) brevi' e 'ossitoni primari', sulla base di quanto detto nei §§2.1.-2.4., devono essere considerate equivalenti.

[28] Le registrazioni, effettuate *in loco*, sono state realizzate in ambienti poco rumorosi, seppur non insonorizzati.

Per l'analisi strumentale è stato usato il *software* Praat (versione 4.6.36; cfr. Boersma & Weenink, 2005). L'estrazione dei parametri è stata possibile grazie ad uno *script* realizzato da Beat Siebenhaar (Università di Lipsia) che permette di misurare anche i valori formantici dei segmenti etichettati. Dopo l'applicazione dello *script* viene creato automaticamente un *file* Log.xls contenente i valori dei dati misurati. In questo modo, una volta etichettato,[29] il *corpus* potrà essere utilizzato per altri tipi di analisi strumentale.

Per l'etichettatura ci si è avvalsi del confronto contemporaneo di forma d'onda e spettrogramma: ogni segmento è stato delineato ed etichettato usando l'alfabeto SAMPA. Ne forniamo un esempio in Fig. 4:

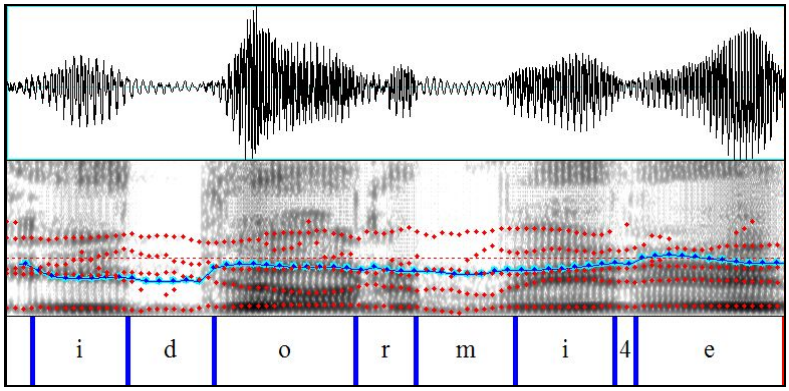


Figura 4: Esempio di segmentazione estratto dalla frase *se potessi dormirei da mia nonna* prodotta dal soggetto SC-M di Castello.

Ovviamente, non sfugge il margine di rischio che un'analisi di questo tipo comporta. La misurazione della durata di una vocale tonica finale di parola, infatti, deve tenere conto di eventuali prolungamenti, che nel parlato spontaneo sono parte integrante e attiva del processo di pianificazione (cfr. Giannini 2003). Dal momento però che nel nostro caso abbiamo fatto ricorso a una forma di parlato elicitato o tutt'al più semispontaneo, e che la brevità delle frasi scelte per l'inchiesta (cfr. la Tab. 3) non ha richiesto ai parlanti particolari sforzi di programmazione, ci sentiamo, almeno in parte, riparati nei confronti di questa criticità.[30] Nel caso poi di palesi enfatizzazioni o di pause sospensive che hanno sensibilmente alterato l'enunciazione della frase, il dato è stato considerato nullo ai fini dell'analisi. Per quanto concerne la misurazione della durata della vocale tonica degli ossitoni in isolamento e in posizione finale di frase, abbiamo tenuto conto di altri parametri, come il decadimento dell'ampiezza.[31]

[29] Per quanto concerne le specifiche della segmentazione, abbiamo cercato di attenerci il più possibile alle tre 'massime' enunciate da Magno Caldognetto (1988: 61): 1) adeguatezza, 2) affidabilità, 3) significatività.

[30] Sul problema della sensibilità alle oscillazioni della struttura informazionale tipica anche del parlato semispontaneo vedi anche Zmarich *et al.* (1997).

[31] Ai fini dell'analisi, abbiamo classificato il contesto di isolamento sempre insieme al contesto finale di enunciato, posizione che risente degli effetti dell'allungamento prosodico (cfr. Marotta, 1985: 100ss.).

Come si vedrà, nei dati raccolti è possibile ravvisare una coerenza che dovrebbe garantire dell'attendibilità delle misurazioni.

3.2. Risultati della prima inchiesta

In questa prima inchiesta l'obiettivo principale della nostra indagine è stato, ovviamente, la misurazione della durata della vocale tonica degli ossitoni presenti nel *corpus*. Nonostante la mancanza di uniformità del materiale raccolto dovuta alla già menzionata situazione dialettale pericolante, la rappresentazione dei dati analizzati ha sostanzialmente confermato quel principio di coerenza territoriale già presente nella distribuzione sia delle isofone (cfr. Tab. 1) sia degli ossitoni secondari (cfr. Tab. 2). Un discorso a parte va fatto per l'unica varietà fuori dal comprensorio della valle del Limentra Occidentale: l'informatore di Lagacci, infatti, ha realizzato gli ossitoni primari e i pochi ossitoni secondari attestati nella sua varietà (cfr. ancora Tab. 2) in maniera assolutamente confusa e sovrapposta riguardo al parametro della lunghezza. L'idiosincratica lentezza d'eloquio ha amplificato questa situazione: tutte le vocali toniche in sillaba libera degli ossitoni superano abbondantemente i 100 ms, e addirittura alcuni ossitoni primari in contesto interno di frase, che ci si sarebbe attesi brevi per concomitanti motivi fonologici e prosodici, hanno riportato vocali di durata superiore ai 200 ms. Dunque, abbiamo ristretto ulteriormente la tabulazione dei dati ai tre informatori di Castello, Cavanna e Stabiazzoni, avendo accertato per il lagaccese, nonostante il frammentato quadro complessivo, l'assenza di qualsiasi opposizione di quantità vocalica in ossitonia.

Abbiamo scelto di rappresentare i dati mediante dei *box-plots* (diagrammi a scatola e baffi), realizzando i grafici con il programma SPSS 10.0 per Windows. I *box-plots* rappresentano una distribuzione statistica di una variabile, come è esemplificato in Fig. 5 (cfr. Tukey, 1977):

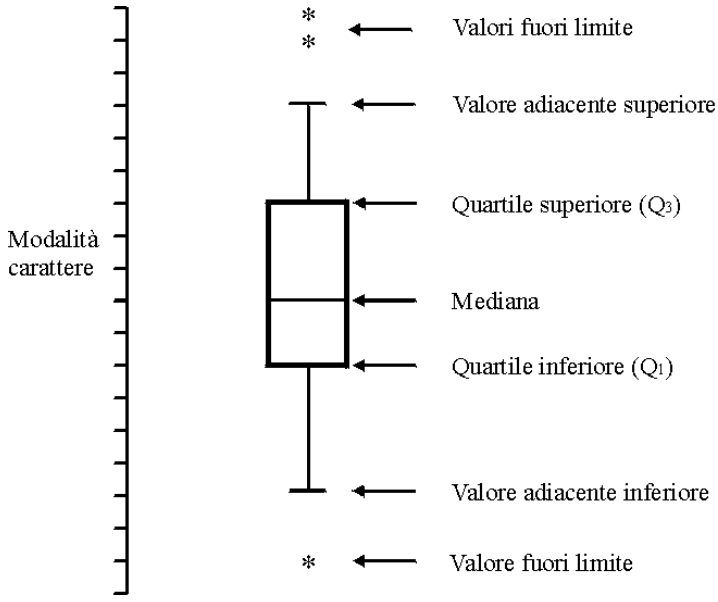


Figura 5: Esempio di rappresentazione statistica di box-plot

Il segmento interno alla scatola rappresenta la 'mediana' della distribuzione, mentre i lati inferiore e superiore della scatola rappresentano rispettivamente il primo ed il terzo quartile. La distanza tra il terzo ed il primo quartile è detta 'distanza interquartilica' e indica la misura della dispersione della distribuzione. Le distanze tra ciascun quartile e la mediana forniscono informazioni relative alla forma della distribuzione: se una distanza è diversa dall'altra, allora la distribuzione è asimmetrica. I segmenti che si allungano dai bordi della scatola ('baffi') individuano gli intervalli in cui sono posizionati i valori rispettivamente inferiori a Q_1 e superiori a Q_3; i punti estremi dei 'baffi' evidenziano i cosiddetti 'valori adiacenti' (cfr. Cleveland, 1993).

I risultati sono presentati per località. In ogni *box-plot* (cfr. Figg. 6-7-8) si confrontano le mediane della durata delle vocali toniche degli ossitoni: da una parte (colonne di sinistra) di quelle (presunte) lunghe nei contesti di isolamento e interno di frase, dall'altra di quelle brevi negli stessi contesti prosodici.

Cominciamo dai dati di Castello: innanzitutto, si può dire che in questa varietà, che, come si è visto, presenta la gamma più ampia di ossitoni secondari tra le varietà qui analizzate, degli ossitoni in sillaba libera con vocale tonica lunga effettivamente esistono.

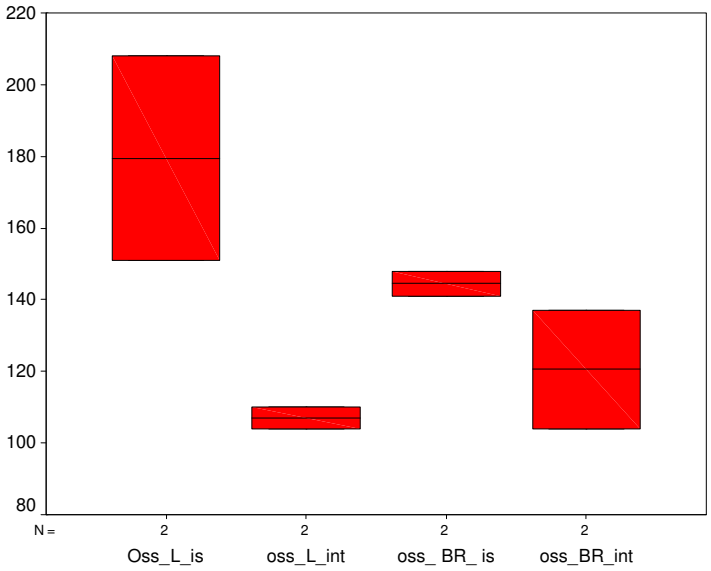


Figura 6: Confronto tra ossitoni lunghi e ossitoni brevi nel contesto isolato e interno per la località di Castello

Il grafico mette in luce come in contesto di isolamento gli ossitoni con vocale tonica lunga presentino una durata sensibilmente maggiore rispetto a quelli con vocale tonica breve (prima e terza colonna): la differenza tra le due mediane è infatti nell'ordine dei 40 ms (180 ms *vs.* 140 ms). Tuttavia, dal confronto dei valori di durata in contesto interno di frase emerge un dato nettamente diverso: la scala in questo contesto interno si inverte e sono gli ossitoni con vocale tonica breve a misurare, seppur di poco, di più degli ossitoni con vocale tonica lunga.

Questo ci induce a pensare che gli ossitoni secondari nella varietà di Castello non vadano considerati lunghi a livello strutturale, ma che la loro lunghezza fonetica sia comunque ammessa, ed emerga nel contesto in cui la realizzazione è più accurata: se si trattasse solo di allungamento prosodico, anche gli ossitoni primari dovrebbero mostrare questo comportamento. D'altra parte, se questa lunghezza avesse valore distintivo, la differenza di durata con tra ossitoni primari e secondari dovrebbe emergere soprattutto nel contesto interno di frase (cfr. Loporcaro *et al.*, 2006: 513-514).

La figura seguente (Fig. 7), concernente Cavanna, delinea una situazione analoga: in contesto di isolamento gli ossitoni secondari presentano una durata notevolmente maggiore rispetto agli ossitoni primari, con le mediane attestate rispettivamente sui 120 ms e sui 70 ms, mentre in contesto interno la situazione si presenta di nuovo appiattita, addirittura leggermente ribaltata. Dunque, tenendo conto di quanto appena argomentato circa i dati di Castello, si può dire che anche nel caso di Cavanna emerge una evidente consistenza fonetica della lunghezza della vocale tonica negli ossitoni secondari, che però non attiva alcuna opposizione di quantità vocalica.

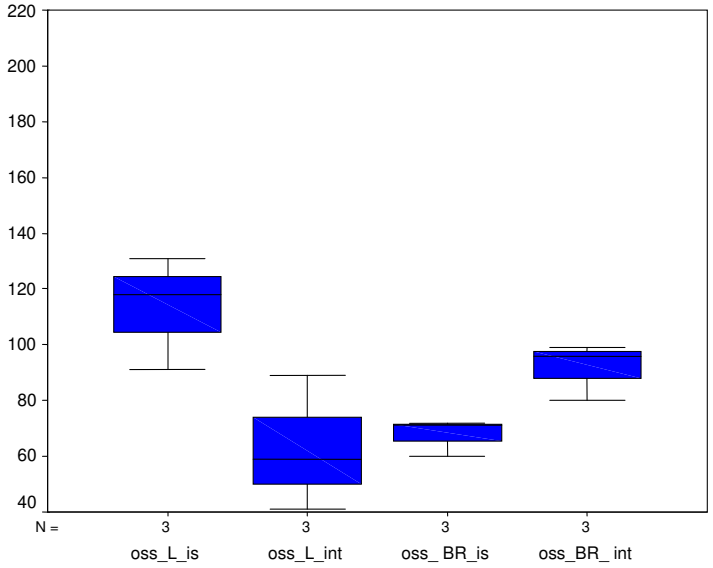


Figura 7: Confronto tra ossitoni lunghi e ossitoni brevi nel contesto isolato e interno per la località di Cavanna

Il quadro di Stabiazzoni (Fig. 8) appare invece completamente appiattito: le mediane delle durate si distribuiscono infatti tutte sulla stessa fascia intorno ai 100 ms.

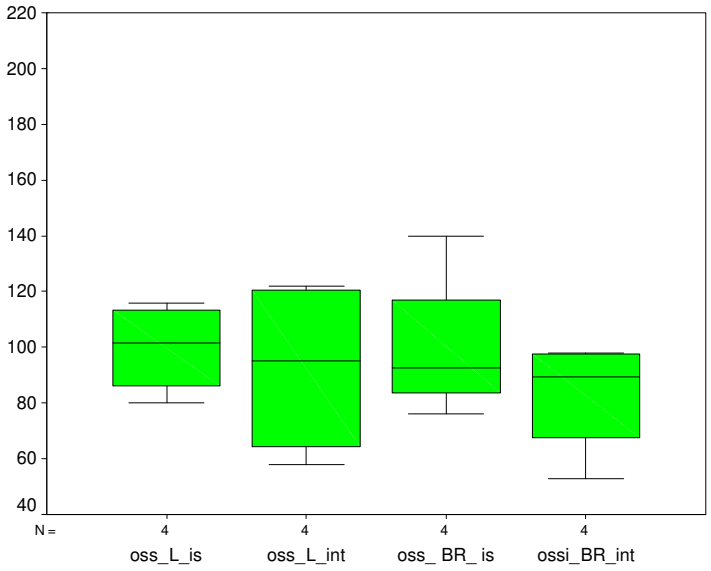


Figura 8: Confronto tra ossitoni lunghi e ossitoni brevi nel contesto isolato e interno per la località di Stabiazzoni

In questo caso né si intravedono gli effetti dell'allungamento prosodico, né si individua una qualche consistenza fonetica della lunghezza della vocale tonica degli ossitoni secondari. Evidentemente, alla luce dei dati a nostra disposizione, la varietà di Stabiazzoni dimostra un trattamento parametrico della durata degli ossitoni, primari e secondari, analoga a quella del toscano. Essendo Stabiazzoni la località più meridionale nella valle del Limentra Occidentale tra quelle analizzate, e dunque la più prossima territorialmente alle varietà pienamente toscane, si può concludere, come già anticipato, che anche questi dati sono conformi a un principio di coerenza territoriale.

3.3. Approfondimento su Castello: risultati della seconda inchiesta

In considerazione di questa situazione piuttosto fluida, abbiamo deciso di svolgere una seconda inchiesta sul campo, concentrandoci sull'informatore di Castello, che, oltre a essere particolarmente attendibile, è estremamente importante in quanto uno degli ultimi testimoni di questa varietà, la meno distante geograficamente e linguisticamente dal pavanese e dal lizzanese nell'ambito di quelle prese in esame questa ricerca. Ferma restando la base dei 57 ossitoni primari e secondari, è stato creato un nuovo *corpus* che permettesse di verificare all'interno di frase i rapporti di durata tra la vocale tonica in sillaba libera degli ossitoni e la consonante iniziale della parola successiva. Inoltre, sono state preparate frasi in cui fosse possibile analizzare a parità di contesti timbrici i rapporti di durata tra vocale tonica e consonante postonica all'interno di parola: a questo scopo, sono state individuate due serie di parossitoni, la prima in cui ci aspettavamo vocale tonica lunga e consonante

postonica scempia, la seconda in cui ci aspettavamo al contrario vocale tonica breve e consonante postonica geminata.[32]

Utilizzando gli stessi strumenti di lavoro elencati nel §3.1. abbiamo prodotto degli istogrammi indicanti le medie delle durate sia delle vocali toniche degli ossitoni primari e secondari e dei parossitoni, sia delle consonanti seguenti (Fig. 9). Per quanto concerne gli ossitoni secondari abbiamo riportato due serie di dati, tenendo conto della possibilità che i numerali *due* e *sei* (rispettivamente [do:] e [se:]) potessero subire degli abbreviamenti in proclisi:[33] essi sono stati inclusi nel calcolo delle medie nella categoria 'ossitoni secondari*', mentre sono stati espunti nella analoga categoria non asteriscata. L'espunzione ha causato, conformemente alle aspettative, un lieve aumento dello iato tra la media della durata vocalica e quella della durata consonantica (prime due colonne di sinistra).

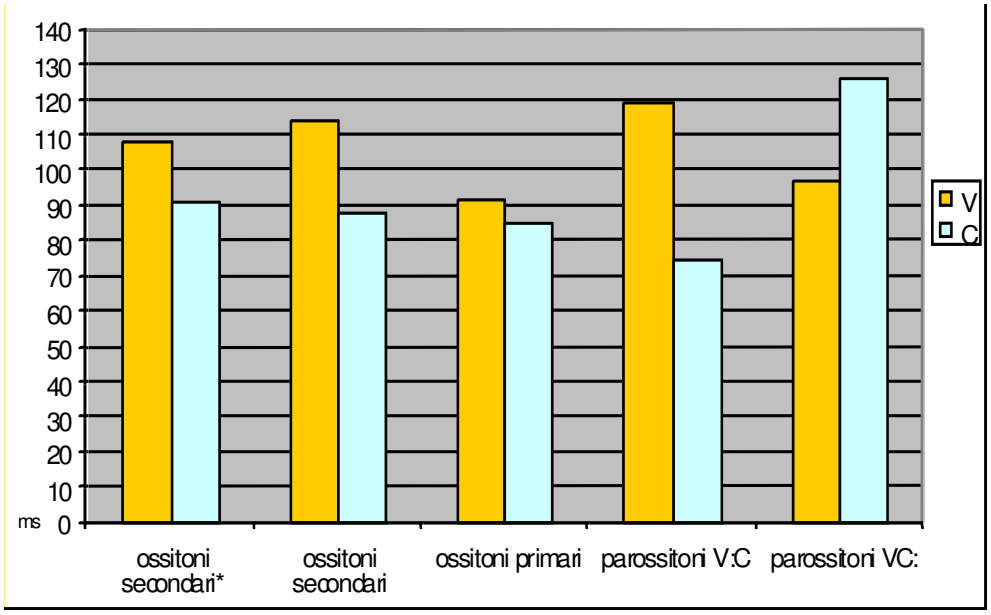


Figura 9: Medie delle durate vocaliche e consonantiche del soggetto SC-M

[32] Per esempio, considerando un contesto timbrico ˈV(#)C come [ˈa(:)(#)p(:)], sono state scelte le seguenti frasi (cfr. i caratteri corsivi): (1) ossitoni secondari – "i farmacisti usavano i mort*ai p*iù piccoli per pestare le erbe"; (2) ossitoni primari – "abbiamo lavorato in citt*à p*er vent'anni"; (3) parossitoni ˈV:C – "lui è il c*ap*o della banda"; (4) parossitoni ˈVC: – "ho tolto il t*app*o della bottiglia". In generale, nella selezione dei contesti timbrici da analizzare, abbiamo privilegiato esempi con consonanti postoniche occlusive.

[33] Ricordiamo a questo punto che in entrambi i *corpora* sono stati aggiunti alla lista di Tab. 2 i pronomi personali *noi*, *voi*, che a Lizzano suonano [nu], [vu] da precedenti [nu:], [vu:] metafonetici abbreviati in proclisi (Malagoli, 1930: §§48, 62); presso alcuni dei nostri informatori abbiamo invece riscontrato l'esito [no:], [vo:], sia in funzione di soggetto, sia in funzione di complemento. Per quanto concerne i possessivi, abbiamo considerato nei *corpora* come ossitoni con vocale tonica potenzialmente lunga soltanto i pronomi (*i*) *miei*, (*i*) *tuoi*, (*i*) *suoi* (cfr. la nota 15).

Concentriamoci ora sui parossitoni (due colonne di destra), la cui situazione risulta piuttosto chiara: osservando in particolare il dato della lunghezza consonantica, si nota come il rapporto tra scempie (colonna V:C) e geminate (colonna VC:) postoniche sia prossimo a 1:2. Questo conferma anche per la varietà di Castello, che, come si è visto (§2.2.), non presenta raddoppiamento fonosintattico e manifesta una generalizzata degeminazione protonica, la presenza ancora salda della geminazione postonica. Abbiamo già visto però che questo dato non è sufficiente a livello diagnostico per la verifica della presenza o meno dell'opposizione di quantità vocalica:[34] bisogna dunque confrontare ossitoni secondari e primari (seconda e terza colonna da sinistra). Innanzitutto, va osservato che i dati raccolti in questa seconda tornata presentano una situazione diversa da quella emersa durante la prima indagine (cfr. la Fig. 6). Anche nel qui rappresentato contesto interno, infatti, la durata media della vocale tonica degli ossitoni secondari è superiore (di 20 ms) a quella della vocale tonica degli ossitoni primari. In presenza di questi ultimi il rapporto tra le durate della vocale tonica (fonologicamente breve) e della consonante postonica (iniziale della parola seguente) tende a 1:1, fatto prevedibile vista l'assenza del raddoppiamento fonosintattico. Ma, benché questo rapporto penda negli ossitoni secondari leggermente a favore della vocale tonica, gli elementi a disposizione, parimenti alla prima tornata di analisi, non ci sembrano sufficienti per decretare che siamo di fronte a una varietà con opposizione di quantità vocalica. Dietro le medie degli ossitoni secondari indicate dagli istogrammi si cela infatti una fortissima variabilità,[35] al cui interno abbiamo rilevato numerosi casi in cui, a fronte di una vocale tonica di durata considerevole, la consonante successiva, subendo un allungamento, si è praticamente appoggiata ad essa, come a ricostruire quel rapporto di durata tendente a 1:1 caratteristico, come abbiamo appena visto, degli ossitoni primari. Siamo dunque ben lontani dalla situazione di netta e costante differenza tra ˈV: e C rilevata nei parossitoni (cfr. ancora Fig. 9, seconda colonna da destra). In altre parole, in un numero considerevole di casi il valore di durata della consonante postonica (iniziale di parola successiva) non mantiene le caratteristiche di una consonante scempia in presenza di vocale lunga nell'ossitono precedente, ma sembra allinearsi all'allungamento di quest'ultima. Questa tendenza all'oscuramento della lunghezza della vocale dell'ossitono può essere considerata come un'ulteriore prova del fatto che la varietà di Castello non abbia quantità vocalica distintiva.[36]

---

[34] Ancora più probante del caso di Lizzano (cfr. la nota 12 sul problema del peso specifico della geminazione postonica) è il già menzionato (§2.3.) caso di Pàvana: la condizione degli ossitoni secondari, analoga a quella del lizzanese, convive con una ancora saldissima presenza di consonanti postoniche geminate. Ma, vista la diagnostica basata sulle indicazioni di Martinet, bisogna postulare per Pàvana presenza di quantità vocalica distintiva.

[35] Anche il dato della deviazione standard conferma questa variabilità: nel caso degli ossitoni primari abbiamo μV = 92 ms (σ = 24); μC = 85 ms (σ = 20,2); μ(V/C) = 1,13 (σ = 0,34); nel caso degli ossitoni secondari μV = 114 ms (σ =29,5); μC = 88 ms (σ = 25,9); μ(V/C) = 1,38 (σ = 0,44). Come si può osservare, σ è per tutti i dati concernenti gli ossitoni secondari nettamente superiore.

[36] Nella già citata (§2.3.) coppia minima del lizzanese la vocale tonica di [kanˈta] presenta una durata di 65 ms, mentre quella di [kanˈta:] arriva a 134 ms (Loporcaro *et al.*, 2006: 513-514). Tale coppia minima, reperibile anche nel pavanese, è invece impossibile a Castello, dove il participio passato non è apocopato (*cantado*, *dormido*, ecc.).

Vediamo due esempi di ossitono con vocale tonica lunga e allungamento fonetico della consonante successiva:

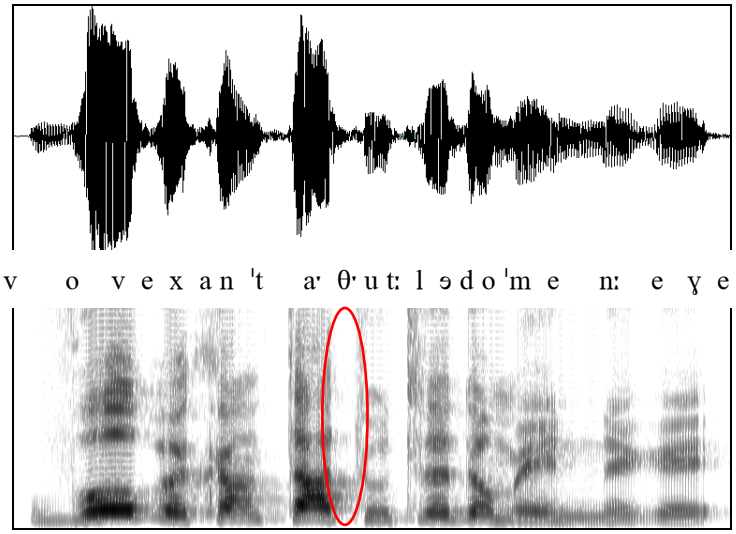


Figura 10: Forma d'onda e spettrogramma della frase
"voi cantate tutte le domeniche" prodotta dal soggetto SC-M {audio 2}

Come si vede dallo spettrogramma in Fig. 10 l'occlusiva dentale /t/ di 'tutte' viene prodotta come fricativa labiodentale [θ], che misura però ben 101 ms, di contro all'occlusiva dentale geminata in [ˈθutːlə] che misura solo 94 ms. Possiamo dunque dire che all'*output* debole rappresentato dalla fricativa [θ] soggiace un segmento strutturalmente breve, che, in presenza di una vocale foneticamente lunga, subisce un allungamento che lascia inalterato il modo di articolazione. Questo conferma il fatto che l'allungamento fonetico della consonante avviene in un contesto fonologico che comunque esclude il rafforzamento fonosintattico "fonologicamente motivato" (Agostiniani, 1992: 4).[37]

Un discorso analogo può essere fatto per lo spettrogramma in Fig. 11:

[37] Secondo Agostiniani (1992: 4) il rafforzamento fonologicamente motivato è "provocato dalla vocale finale di parola quando accentata". È lo stesso Agostiniani (1992: 7-8) a richiamare l'attenzione sulla condizione di occlusione piena delle consonanti sotto raddoppiamento fonosintattico in toscano: in caso di assenza di questa condizione, il segmento va considerato come strutturalmente debole, e l'eventuale allungamento, come nei casi che qui stiamo discutendo, deve essere ascritto a ragioni allofoniche. Avevamo già osservato (§2.2.) il fenomeno, tipico dei dialetti di Sambuca Pistoiese, della sovraestensione della gorgia toscana in vece del raddoppiamento fonosintattico e come succedaneo della degeminazione protonica. Agostiniani (1992: 2) ricorda peraltro che la conservazione della lunghezza distintiva delle consonanti è una condizione necessaria, ma non sufficiente, per la presenza di raddoppiamento fonosintattico. Il dialetto di Castello inquadra perfettamente questa casistica: infatti, nonostante il raddoppiamento fonosintattico sia assente, la conservazione della lunghezza distintiva delle consonanti risulta conservata.

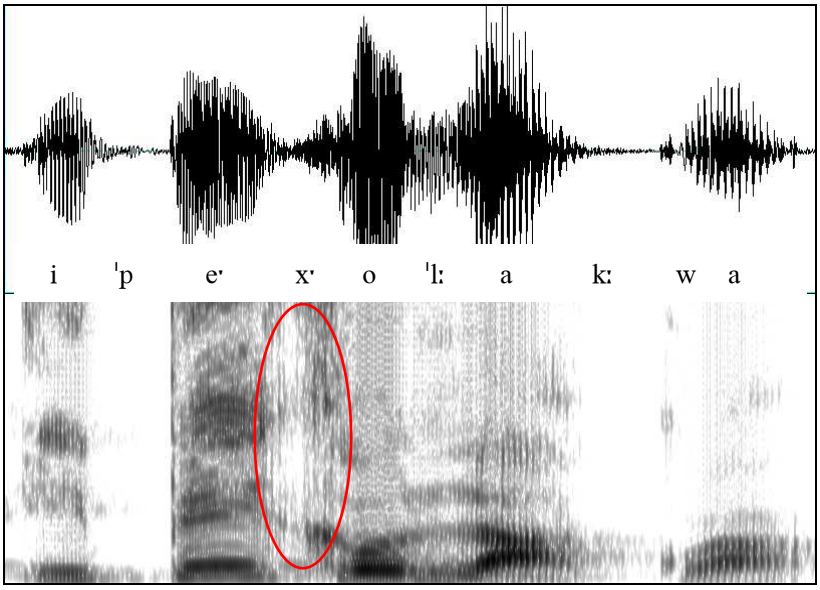


Figura 11: Forma d'onda e spettrogramma della frase (voi vi lavate) "i piedi con l'acqua" prodotta dal soggetto SC-M {audio 3}

Anche in questo caso l'*output* fonetico debole [x] della /k/ di *con* è garanzia della soggiacenza di un segmento fonologicamente breve. Il risultato fonetico, però, è quello di una spirante velare di durata (120 ms) pressoché analoga a quella (118 ms) della vocale tonica dell'ossitono precedente [peˑ]. Anche in questo caso, quindi, l'instaurarsi del rapporto 1:1 tra vocale tonica e consonante successiva sembra teso a obliterare la lunghezza della prima, quasi a volerne ribadire la non rilevanza fonologica.

4. CONCLUSIONI

Alla luce di quanto presentato e discusso, si possono trarre alcune conclusioni. Innanzitutto, tra le quattro località analizzate, Lagacci e Stabiazzoni (in particolare quest'ultima) sembrano restituire condizioni pienamente toscane per quanto concerne il trattamento degli ossitoni. Sia gli ossitoni primari, sia gli ossitoni secondari formatisi in queste varietà (cfr. Tab. 2), sono parametricamente brevi: i dati che abbiamo presentato per Stabiazzoni (Fig. 6) indicano che questa brevità è fatto non solo strutturale, ma anche fonetico, in tutti i contesti.

Cavanna e Castello, invece, presentano in posizione di isolamento una distinzione piuttosto netta tra le durate delle vocali toniche degli ossitoni primari (brevi) e secondari (lunghe). Questa distinzione sparisce completamente in contesto interno di frase a Cavanna, mentre i dati di Castello si presentano più disomogenei. Al netto di eventuali allungamenti prosodici o di possibili prolungamenti vocalici dovuti a programmazione del parlato (§3.1.), sembra che, in ogni caso, ossitoni in sillaba libera con vocale tonica lunga siano ammessi in queste due varietà (§ 2.2.). Abbiamo assistito, però, a strategie di riparazione in contesto interno operate mediante l'allungamento del segmento consonantico che segue la vocale

tonica (§ 2.3.). Viceversa, la consistenza della lunghezza consonantica in postonia nei parossitoni appare sostanzialmente intatta (§ 2.3.).

Il quadro che se ne deduce è complesso ma diacronicamente, diatopicamente e strutturalmente coerente. Prendendo a modello la varietà di Castello, possiamo ipotizzare che il verificarsi di condizioni atte all'insorgere di ossitoni secondari (la caduta di *-i* in iato primario e secondario) abbia creato un gruppo di parole con una vocale tonica la cui effettiva lunghezza fonetica è stata considerata strutturalmente irrilevante, come dimostra il frequente allungamento consonantico in contesto interno. Quindi, se da una parte è saltato il parametro toscano della brevità degli ossitoni in sillaba libera, dall'altra non si è determinato il passaggio delle consegne della lunghezza distintiva tra consonanti e vocali, come è accaduto, poco più a nord, a Pàvana e a Lizzano. Piuttosto, la scomparsa del parametro della brevità ha determinato una sorta di rottura dei ranghi tra ossitoni primari e secondari: i dati che abbiamo raccolto nella nostra seconda tornata di indagini riportano per il contesto di isolamento una durata media della vocale tonica negli ossitoni primari pari a ben 159 ms, di contro a una di 182 ms negli ossitoni secondari.

Siamo di fronte a una tipica fenomenologia da area grigia, che potrebbe essere messa in parallelo con quadri consimili, come quello prospettato per l'italiano di Carrara da Barbera (in stampa). Là si erano registrate, seppur disomogenee negli informatori, geminazioni spurie in parossitoni con vocale tonica lunga (in particolare nei participi passati in *-ato -ito -uto*), messe in relazione dall'autore con l'omogenea (questa sì) degeminazione protonica, in un quadro di sostanziale indebolimento del parametro fonologico di quantità. Nel nostro caso le geminazioni spurie riguardano i contesti ossitoni: di fronte alla saldezza dei rapporti di durata nei parossitoni, l'indebolimento del parametro fonologico di quantità interessa qui proprio quella posizione da cui è possibile, seguendo Martinet, diagnosticare la presenza di lunghezza vocalica distintiva. La nostra diagnosi, che non sarebbe stata possibile se ci si fosse limitati alla misurazione della durata vocalica, mostra quindi dal vivo una situazione di transizione in cui l'ossitonia manifesta un punto di criticità del sistema, pronto a quella ristrutturazione fonologica che caratterizza ancor oggi molti dialetti a nord della linea La Spezia-Rimini (o Carrara-Fano), a cavallo della quale è situata Sambuca Pistoiese.

RINGRAZIAMENTI

I due Autori desiderano ringraziare l'Archivio Fonografico dell'Università di Zurigo per la fornitura degli strumenti utilizzati per le registrazioni; Beat Siebenhaar per avere concesso l'utilizzo dello script di *Praat* da lui creato; Arianna Uguzzoni, Michele Loporcaro e Vincenzo Faraoni per i preziosi suggerimenti, esimendoli ovviamente dalla responsabilità degli errori e delle imperfezioni presenti nel testo.

5. BIBLIOGRAFIA

Agostiniani, L. (1989), Fenomenologia dell'elisione nel parlato in Toscana, *Rivista Italiana di Dialettologia*, 13, 7-46.

Agostiniani, L. (1992), Su alcuni aspetti del 'rafforzamento sintattico' in Toscana e sulla loro importanza per la qualificazione del fenomeno in generale, *Quaderni del Dipartimento di Linguistica dell'Università di Firenze*, 3, 1-28.

AIS = Jaberg, K. & Jud, J. (1928-40), *Sprach- und Sachatlas Italiens und der Südschweiz*, Zofingen: Ringier.

Barbato, M. (in stampa), *Dio mio*. Un frammento di grammatica storica, in *Actes du XXV Congrès de Linguistique et Philologie Romanes*, Innsbruck, 3-8 settembre 2007.

Barbera, M. (in stampa), Toscani a metà. Degeminazione e geminazione nell'italiano di Carrara, in *Kontaminationen – Contaminations – Contaminazioni – Contaminaciones. Atti del IV Dies Romanicus Turicensis*, Zurigo, 16-17 novembre 2007.

Boersma, P. & Weenink, D. (2005), *PRAAT: doing phonetics by computer*, www.fon.hum.uva.nl/praat/.

Carpitelli, E. & Iannaccaro, G. (1995), Dall'impressione al metodo: per una ridefinizione del metodo escussivo, in *Dialetti e lingue nazionali. Atti del XXVII Convegno della S.L.I.* (M.T. Romanello & I. Tempesta, editors), Roma: Bulzoni, 99-120.

Castellani, A. (1952), *Nuovi testi fiorentini del Dugento*, *Tomo I*, Firenze: Sansoni.

Cleveland, W.S. (1993), *Visualizing data*, Murray Hill: AT&T Laboratories.

Coco, F. (1970), *Il dialetto di Bologna. Fonetica storica e analisi strutturale*, Bologna: Forni.

DEI = Battisti, C. & Alessio, G. (1950-7), *Dizionario Etimologico Italiano*, Firenze: Barbera.

Filipponio, L. (2007a), Alcuni dati sul trattamento dei proparossitoni etimologici nei dialetti dell'Appennino Bolognese, in *Scienze vocali e del linguaggio. Metodologie di valutazione e risorse linguistiche* (V. Giordani, V. Bruseghini & P. Cosi, editors), Atti del 3° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, 29 novembre – 1° dicembre 2006, Povo (Trento), Torriana (RN): EDK editore, 91-100.

Filipponio, L. (2007b), *Lingua e storia nei dialetti della valle del Reno*, Porretta Terme: Gruppo di Studi Alta Valle del Reno – Nuèter.

Filipponio, L. (2007c), Le cose, le parole, il dialetto, in *Torri: Museo della vita quotidiana. Collezione Renzo Innocenti* (P. Gioffredi, editor), San Giovanni Valdarno: Industria Grafica Valdarnese, 21-23.

Filipponio, L. (2008a), I Liguri a Treppio: breve storia di un fraintendimento, *Nuèter*, 67, 128-132.

Filipponio, L. (2008b), La guarzetta vien dalla montagna, *Nuèter*, 68, 130-137.

Filipponio, L. (in corso di stampa a), *La struttura di parola dei dialetti della valle del Reno*, Sala Bolognese: Forni.

Filipponio, L. (in corso di stampa b), La quantità vocalica nei proparossitoni etimologici al confine tra toscano e gallo-italico, in *Actes du XXV Congrès de Linguistique et Philologie Romanes*, Innsbruck, 3-8 settembre 2007.

Giannini, A. (2003), Prolungamenti vocalici e vocalizzazioni, in *Voce, canto, parlato. Studi in onore di Franco Ferrero* (E. Magno Caldognetto, P. Cosi, A. Zamboni, editors), Padova: Unipress: 163-172.

Guccini, F. (1998), *Dizionario del dialetto di Pàvana*, Porretta Terme: Gruppo di Studi Alta Valle del Reno – Nuèter.

Haudricourt, A.G. & Juilland, A.G. (1949), *Essai pour une histoire structurale du phonétisme français*, Paris: Klincksieck.

Jaberg, K. & Jud, J. (1987), *L'atlante linguistico come strumento di ricerca. Vol. 1: Fondamenti critici e introduzione*, Milano: Unicopli (Trad. it. di S. Baggio a cura di G. Sanga di *Der Sprachatlas als Forschungsinstrument: kritische Grundlegung und Einführung in den Sprach- und Sachatlas Italiens und der Südschweiz*, Halle: Niemeyer, 1928).

Loporcaro, M. (1997), *L'origine del raddoppiamento fonosintattico*, Basel-Tübingen: Francke.

Loporcaro, M. (2005), La lunghezza vocalica nell'Italia settentrionale alla luce dei dati del lombardo alpino, in *Itinerari linguistici alpini. Atti del convegno di dialettologia in onore del prof. Remo Bracchi* (M. Pfister & G. Antonioli, editors), Sondrio: Istituto di Dialettologia e di Etnografia Valtellinese e Valchiavennasca – LEI, 97-113.

Loporcaro, M. (2005-6), I dialetti dell'Appennino Tosco-emiliano e il destino delle atone finali nel(l'italo-)romanzo settentrionale, *L'Italia Dialettale*, 66-67, 69-122.

Loporcaro, M., Paciaroni, T. & Schmid, S. (2005), Consonanti geminate in un dialetto lombardo alpino, in *Misura dei parametri. Aspetti tecnologici ed implicazioni nei modelli linguistici* (P. Cosi, editor), Atti del 1° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Padova, 2-4 dicembre 2004, Torriana (RN): EDK Editore, 597-618.

Loporcaro, M., Delucchi, R., Nocchi, N., Paciaroni, T. & Schmid, S. (2006), La durata consonantica nel dialetto di Lizzano in Belvedere (Bologna), in *Analisi prosodica. Teorie, modelli e sistemi di annotazione* (R. Savy & C. Crocco, editors), Atti del 2° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Salerno, 30 novembre – 2 dicembre 2005, Torriana (RN): EDK Editore, 491-517.

Lüdtke, H. (1956), *Die strukturelle Entwicklung des romanischen Vokalismus*, Bonn: Romanisches Seminar an der Universität Bonn.

Magno Caldognetto, E. (1988), L'apporto delle tecniche sperimentali alla descrizione fonetica: alcuni esempi, in *Guida ai dialetti veneti* (M. Cortelazzo, editor), Padova: CLUEB, 10, 61-83.

Malagoli, G. (1930), Fonologia del dialetto di Lizzano in Belvedere (Appennino bolognese), *L'Italia Dialettale*, 6, 125-196.

Malagoli, G. (1940), Appunti di morfologia e sintassi del dialetto di Lizzano in Belvedere, *L'Italia Dialettale*, 16, 191-211.

Marotta, G. (1985), *Modelli e misure ritmiche: la durata vocalica in italiano*, Bologna: Zanichelli.

Martinet, A. (1955), *Economie des changements phonétiques. Traité de phonologie diachronique*, Bern: Francke.

Martinet, A. (1966), Close contact, *Word*, 22, 1-6.

Martinet A. (1975), *Evolution des langues et reconstruction*, Paris: Presses Universitaires de France.

Pellegrini, G.B. (1992), Il 'Cisalpino' e l'italo-romanzo, *Archivio Glottologico Italiano*, 77, 272-296.

Richter, E. (1934), *Beiträge zur Geschichte der Romanismen, I. Chronologische Phonetik des Französischen bis zum Ende des 8. Jahrhunderts*, Halle: Max Niemeyer.

Rohlfs, G. (1966), *Grammatica storica della lingua italiana e dei suoi dialetti. Fonetica*, Torino: Einaudi.

Rohlfs, G. (1968), *Grammatica storica della lingua italiana e dei suoi dialetti. Morfologia*, Torino: Einaudi.

Rohlfs, G. (1969), *Grammatica storica della lingua italiana e dei suoi dialetti. Sintassi e formazione delle parole*, Torino: Einaudi.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Reading: Addison-Wesley.

Uguzzoni, A. (1974), Sulla struttura della parola dei dialetti emiliani: aspetti sincronici e aspetti diacronici di un problema, *Deputazione di Storia Patria per le Antiche Provincie Modenesi*, 38, 239-252.

Uguzzoni, A. (1975), Appunti sulla evoluzione del sistema vocalico di un dialetto frignanese, *L'Italia Dialettale*, 38, 47-76.

Uguzzoni, A., Azzaro, G. & Schmid, S. (2003), Short vs long and/or abruptly vs smoothly cut vowels. New perspectives on a debated question, in *Proceedings of the 15t^h International Congress of Phonetic Sciences* (M.J. Solé *et al.*, editors), Barcelona: Spain, 2717-2720.

Ungarelli, G. (1901), *Vocabolario del dialetto bolognese*, Bologna: Tip. Zamorani e Albertazzi.

Vennemann, T. (1988), *Preference laws for syllable structure and the explanation of sound change*, Berlin: Mouton de Gruyter.

Vennemann, T. (2000), From quantity to syllable cuts: on so-called lengthening in the Germanic languages, *Rivista di Linguistica/Italian Journal of Linguistics*, 12, 251-282.

Vitali, D. (2007), Il dialetto di Porretta Terme, *Nuèter*, 65, 52-58.

Vitali, D. (2009), Le guarzette, Torri, il Frignano e Porretta, *Nuèter*, 69, 33-38.

Weinrich, H. (1958), *Phonologische Studien zur Romanischen Sprachgeschichte*, Münster: Aschendorff.

Zmarich, C., Magno Caldognetto, E. & Ferrero, F. (1997), Analisi confrontativa di parlato spontaneo e letto: fenomeni macroprosodici e indici di fluenza, in *Fonetica e fonologia degli stili dell'italiano parlato* (F. Cutugno, editor), Atti delle VII Giornate di Studio del Gruppo di Fonetica Sperimentale, Napoli, 14-15 novembre 1996, Roma: Esagrafica, 111-139.

ELISIONE OBBLIGATORIA, VARIABILE E POCO FREQUENTE NEL FIORENTINO: UN CASO DI ALLOMORFIA FRASALE PRECOMPILATA CON FORME PREFERENZIALI

Luigia Garrapa
Università del Salento & Universität Würzburg
luigiagarrapa@yahoo.it

1. SOMMARIO

Questo lavoro intende esplorare e chiarire l'applicazione dell'elisione nella varietà di italiano parlata a Firenze, individuando i fattori che la condizionano. I contesti presi in esame sono costituiti dai determinanti (*un(o)/una, l(o)/la, quell(o)/quella, questo/questa, le, quelle* e *queste/questi*) seguiti da sostantivi che iniziano per vocale e dai proclitici (*lo/la, li/le*$_{acc}$, *mi/ti, ci/vi* e *le*$_{dat}$) seguiti da verbi lessicali che iniziano per vocale.

I dati analizzati provengono da due fonti: il corpus *C-Oral-Rom* (Cresti & Moneglia, 2005) e le inchieste sul campo condotte dall'Autrice a Firenze in dicembre 2007 e sono rappresentativi del parlato spontaneo e di quello elicitato. In questa sede si discutono i dati congiunti del corpus e delle inchieste e si avanza una proposta per la rappresentazione dell'elisione.

L'elisione nel fiorentino risulta condizionata dalla tipologia delle parole funzionali, dal tratto morfologico di numero e dallo stile discorsivo. In primo luogo, le vocali atone finali dei determinanti subiscono l'elisione con maggiore frequenza rispetto alle vocali finali dei proclitici. In secondo luogo, le vocali finali dei determinanti singolari, dei clitici accusativi singolari e dei clitici di persona vengono elise più frequentemente rispetto alle vocali finali dei determinanti plurali e dei clitici accusativi plurali. In terzo luogo, l'elisione viene applicata con maggiore probabilità nel parlato informale rispetto a quello formale.

La proposta avanzata in questo lavoro è che l'elisione non sia il risultato di un 'vero' processo di cancellazione vocalica, ma che sia le forme terminanti per vocale che le forme elise delle parole funzionali analizzate siano elencate individualmente nel lessico mentale assieme al contesto in cui possono apparire ed alle eventuali preferenze di selezione (fra le forme elise e quelle terminanti per vocale) in contesto prevocalico.

2. INTRODUZIONE

Questo studio si propone di chiarire il funzionamento dell'elisione delle vocali atone finali dei determinanti e dei proclitici, identificando i fattori che ne favoriscono l'applicazione.

Precedenti studi sull'elisione in diverse varietà di italiano hanno affermato che, eccezion fatta per l'elisione obbligatoria con i determinanti *un(o), l(o)* e *quell(o),* l'elisione è estremamente variabile con i restanti determinanti singolari e plurali e con tutti i proclitici (Vogel *et al.,* 1983; Agostiniani, 1989; Marotta & Sorianello, 1997; Marotta, 1995; Nespor, 1990). Più in particolare, questi studi mettono in evidenza che l'elisione viene in parte condizionata dal tratto morfologico di numero, che risulta in qualche modo impedita dalla presenza di un accento di parola sulla vocale successiva, e che è soggetta ad estrema variazione *interspeaker* ed *intraspeaker* tanto da risultare spesso impredicibile. Tuttavia gli studi appena citati si concentrano prevalentemente sull'elisione nelle sequenze di due

parole funzionali e non offrono un'analisi sistematica dell'elisione nei determinanti e nei proclitici. Inoltre, nei suddetti studi mancano precisi riferimenti a dati quantitativi.

Questo lavoro intende fornire un'analisi esaustiva dell'elisione in tutti i determinanti ed i proclitici, mettendo in luce i fattori che ne determinano l'applicazione. La varietà di italiano analizzata è quella parlata a Firenze, nella quale, oltre all'elisione, sono attivi diversi processi di cancellazione vocalica come il troncamento, l'aferesi e l'apocope (Agostiniani, 1989; Marotta, 1995; Marotta & Sorianello, 1997). Seguendo la metodologia tradizionalmente indicata in letteratura (Vaux & Cooper, 1999; Newmann & Ratliff, 2001) ed in precedenti studi sull'elisione (Cabré & Prieto, 2005; Dehé, 2008), questo studio si concentra sull'elisione nel parlato spontaneo ed in quello elicitato e presenta i risultati congiunti per i due livelli di parlato.

In particolare, questo studio intende mettere in luce che l'elisione non è interamente opzionale ed impredicibile e che risulta influenzata dalla categoria della parole funzionali, dal tratto morfologico di numero (ma non da quello di genere) e dallo stile discorsivo. Al contrario, l'accento risulta irrilevante per l'elisione.

Sulla scia di Hayes (1990), Mascaró (2007) e Bonet *et al.* (2007), l'elisione nel fiorentino viene rappresentata come un processo di allomorfia frasale precompilata (talvolta *categorica* e talvolta *graduale*) in cui, per tutte le parole funzionali analizzate, sia le forme terminanti per vocali che quelle terminanti per consonante sono elencate nel lessico mentale assieme alle preferenze selettive delle stesse in contesto prevocalico.

Questo lavoro è strutturato in sette sezioni. Nella sezione 3 vengono presentate le parole funzionali analizzate e la loro composizione morfologica. La sezione 4 è dedicata alla metodologia utilizzata ed alle ipotesi di analisi. La sezione 5 presenta i risultati congiunti del corpus e delle inchieste sul campo, prendendo in considerazione prima i fattori che condizionano l'elisione (la tipologia delle parole funzionali, la morfologia e lo stile discorsivo) e poi quelli che, invece, risultano irrilevanti (l'accento). La sezione 6 discute i risultati ed avanza una proposta per la rappresentazione dell'elisione nel fiorentino. La sezione 7 offre la conclusione.

3. LE PAROLE FUNZIONALI ANALIZZATE E LA SOTTOSPECIFICAZIONE MORFOLOGICA

Le parole funzionali analizzate nel presente lavoro sono i 27 determinanti ed i 9 proclitici presentati nelle Tabelle 1 e 2:

	Sg.		Pl.	
	Masc.	Fem.	Masc.	Fem.
Articoli definiti	[l(o)]	[l(a)]		[l(e)]
Articoli indefiniti	[un(o)]	[un(a)]		
Aggettivi	[kwel:(o)]	[kwel:(a)]		[kwel:(e)]
dimostrativi	[kwest(o)]	[kwest(a)]	[kwest(i)]	[kwest(e)]
Preposizioni	[al:(o)]	[al:(a)]		[al:(e)]
articolate	[dal:(o)]	[dal:(a)]		[dal:(e)]
	[del:(o)]	[del:(a)]		[del:(e)]
	[nel:(o)]	[nel:(a)]		[nel:(e)]
	[sul:(o)]	[sul:(a)]		[sul:(e)]

Tabella 1: I determinanti analizzati

Caso	Persona	Sg.		Pl.	
		Masc.	Fem.	Masc.	Fem.
Acc.	3	[l(o)]	[l(a)]	[l(i)]	[l(e)]
Acc. & Dat.	1	[m(i)]		[tʃ(i)]	
	2	[t(i)]		[v(i)]	
Dat.	3	[le]			

Tabella 2: I proclitici analizzati

Gli articoli definiti *l(o), la* e *le* e tutti i proclitici sono monosillabi, mentre gli articoli indefiniti *un(o)*/*una*, gli aggettivi dimostrativi *questo/quello, questa/quella, questi* e *queste/ quelle* e tutte le preposizioni articolate constano di due sillabe.

Si deve precisare che le parole funzionali elencate in (1) non sono state tenute in con - siderazione:

(1) a. [il] [kwel] [al] [dal] [del] [nel] [sul]
b. [i] [kwei] [ai] [dai] [dei] [nei] [sui]
c. [ʎi] [kweʎ:i] [aʎ:i] [daʎ:i] [deʎ:i] [neʎ:i] [suʎ:i]
d. [ʎi] (clitico dativo maschile singolare)

I determinanti maschili singolari indicati in (1a) ed (1b) sono stati tralasciati per il fatto che essi precedono sostantivi maschili singolari e plurali che iniziano per consonante e, quindi, l'elisione non può avere luogo. La ragione che ha determinato l'esclusione dei determinanti maschili plurali menzionati in (1c) e del clitico dativo indicato in (1d) risiede nel fatto che, sebbene essi precedano sostantivi e verbi che iniziano per vocale, la loro -/i/ finale è preceduta dalla consonante laterale palatale /ʎ/. In effetti, per i parlanti dell'italiano è alquanto difficile, se non impossibile, discriminare l'avvenuta o la mancata elisione di -/i/ preceduta da /ʎ/ basandosi esclusivamente sull'analisi uditiva (cfr. Garrapa, 2009: 115-117).

Non tutte le parole funzionali indicate nelle Tabelle 1 e 2 sono state analizzate sia nel parlato spontaneo che in quello elicitato. Per quanto riguarda i determinanti, tutti gli articoli e gli aggettivi dimostrativi di vicinanza sono stati esaminati nei due livelli di parlato. I dimostrativi di lontananza e tutte le preposizioni articolate, invece, sono stati studiati solo nel parlato spontaneo. Per quanto concerne i proclitici, tutti i clitici accusativi ed i clitici di 1ª persona sono stati studiati sia nel parlato spontaneo che in quello elicitato. I clitici di 2ª persona ed il clitico dativo *le,* invece, sono stati analizzati solo nel parlato informale.

In ciò che segue, si considera la struttura morfologica dei determinanti e proclitici presentati nelle Tabelle 1 e 2. Più in particolare, si considera che queste parole funzionali siano ulteriormente segmentabili in una radice ed una vocale che è generalmente (ma non sistematicamente) un suffisso flessivo. Le categorie morfologiche rilevanti per l'elisione sono quelle di genere e di numero, dal momento che queste categorie sono generalmente espresse dalle vocali bersaglio di elisione. Le categorie morfologiche di persona e di caso, invece, vengono realizzate dalle radici delle parole funzionali in questione. In questo studio, si distingue fra i valori di *default* e quelli marcati propri delle categorie morfologiche di genere, numero, persona e caso. Sebbene i concetti di marcatezza e non marcatezza siano piuttosto controversi (Haspelmath, 2006), sulla scia di Greenberg (1966a-b) si considerano

come valori di *default* (o non marcati) quei valori che ricorrono più frequentemente nel parlato quotidiano ed in prospettiva interlinguistica e/o che sono codificati esplicitamente. I restanti valori, invece, vengono considerati come marcati, si veda la Tabella 3.

Categorie morfologiche	Valori	
	Default	Marcati
Caso	[nominativo]	[accusativo], [dativo]
Persona	[3ª]	[1ª], [2ª]
Numero	[singolare]	[plurale
Genere	[maschile]	[femminile]

Tabella 3: Valori di *default* vs. valori marcati per ciascuna categoria morfologica

Seguendo l'approccio della sottospecificazione morfologica nel lessico mentale (Farkas, 1990; Lahiri & Reetz, 2002; Embick & Noyer, 2005), si assume che i morfemi che realizzano i valori di *default* siano sottospecificati (indicati mediante [ ]) per le corrispondenti categorie morfologiche. Quei morfemi che realizzano i valori marcati, invece, vengono considerati come interamente specificati per i tratti morfologici corrispondenti all'interno del lessico mentale.

Si considera ora la struttura morfologica dei determinanti e proclitici menzionati nelle Tabelle 1 e 2. Si propone che i determinanti siano formati da una radice (che termina in consonante o che consta di un'unica consonante) e da un suffisso flessivo (la vocale finale): la radice è sottospecificata per la categoria morfologica di persona, mentre il suffisso flessivo realizza i tratti di numero ([ ] vs. [plurale]) e genere ([ ] vs. [femminile]); cfr. la Tabella 4:

	Valori morfologici espressi da		
	Radice	Vocale finale	
Determinanti	*Persona*	*Numero*	*Genere*
un-**o,** l-**o,** quell-**o,** quest-**o**	[ ]	[ ]	[ ]
un-**a,** l-**a,** quell-**a,** quest-**a**			[femminile]
quest-**i**		[plurale]	[ ]
l-**e,** quell-**e,** quest-**e**			[femminile]

Tabella 4: Specificazioni morfologiche dei determinanti.

La struttura morfologica dei clitici accusativi è molto simile a quella dei determinanti. In effetti, si propone che i clitici accusativi siano segmentabili in una radice ed un suffisso flessivo. La radice esprime il caso [accusativo] ma è sottospecificata per il tratto di persona. Le vocali finali /o, a, i, e/, invece, realizzano i tratti di numero ([ ] vs. [plurale]) e genere ([ ] vs. [femminile]); cfr. la Tabella 5:

	Valori morfologici espressi da			
	Radice		Vocale finale	
Clitici accusativi	*Caso*	*Persona*	*Numero*	*Genere*
l-**o**	[accusativo]	[ ]	[ ]	[ ]
l-**a**				[femminile]
l-**i**			[plurale]	[ ]
l-**e**				[femminile]

Tabella 5**:** Specificazioni morfologiche dei clitici accusativi

La struttura morfologica dei clitici di persona, invece, è differente rispetto a quella dei clitici accusativi. Sulla scia di Kayne (2000: 134-135) e Cardinaletti & Shlonsky (2004: 532-534), si assume che i clitici di persona non siano ulteriormente segmentabili in una radice consonantica ed in un suffisso flessivo e che, quindi, la /i/ di *mi/ti* e *ci/vi* e la /e/ del clitico dativo *le* non siano suffissi flessivi; si veda la Tabella 6:

Clitici di persona	Valori morfologici realizzati dall'unico morfo			
	Caso	*Persona*	*Numero*	*Genere*
m**i**	[ ]	[1ª]	[ ]	[ ]
t**i**		[2ª]		
c**i**		[1ª]	[plurale]	
v**i**		[2ª]		
l**e**	[dativo]	[ ]	[ ]	[femminile]

Tabella 6: Specificazioni morfologiche dei clitici di persona

Come si vedrà a seguire, l'elisione risulta determinata in gran parte dalla struttura morfologica e soprattutto dal tratto morfologico di numero espresso dalle vocali bersaglio di elisione. Dunque l'elisione si presenta come un fenomeno che coinvolge l'interfaccia morfologia-fonologia, dal momento che le vocali bersaglio di elisione sono in gran parte suffissi flessivi sul piano morfologico.

4. IL METODO, I DATI, L'ANALISI

I dati discussi nel presente lavoro provengono da due fonti: il corpus *C-Oral-Rom* (Cresti & Moneglia, 2005) ed i dati elicitati durante le inchieste sul campo condotte a Firenze in dicembre 2007. I dati del corpus sono rappresentativi del parlato spontaneo, mentre quelli raccolti sul campo sono rappresentativi del parlato elicitato.

4.1 L'analisi del corpus

Il corpus italiano all'interno del corpus romanzo *C-Oral-Rom* (Cresti & Moneglia, 2005) è molto vasto ed è rappresentativo del parlato spontaneo (sia formale che informale) e soprattutto della varietà parlata a Firenze. Il corpus si basa sul parlato di 451 informatori prevalentemente toscani, di diversa età e con diversi livelli di scolarizzazione.

I contesti analizzati sono 4.450. I tre fattori indagati nel corpus sono menzionati in (2):

(2) a. La classe delle parole funzionali (determinanti vs. proclitici);
b. I tratti morfologici di genere ([ ] vs. [femminile]) e di numero ([ ] vs. [plurale]) realizzati dalle vocali bersaglio di elisione;
c. Lo stile discorsivo (parlato informale vs. formale).

Le ipotesi di analisi sono le seguenti: a) le vocali dei determinanti vengono elise più di frequente rispetto a quelle dei proclitici; b) le vocali che sono sottospecificate per il numero e/o per il genere vengono elise con maggiore probabilità rispetto a quelle che sono esponenti morfologici del genere [femminile] ed del numero [plurale]; c) l'elisione avviene più frequentemente nel parlato informale che in quello formale.

I casi di applicazione e di mancata applicazione dell'elisione sono stati individuati avvalendosi di due metodologie. In un primo momento, è stata effettuata l'analisi uditiva su un campione dei *files* audio al fine di appurare l'affidabilità delle trascrizioni dei *files* audio. In un secondo momento, ci si è serviti del software *Contextes,* il quale permette di accedere alle trascrizioni dei *files* audio del corpus senza dover ascoltare gli stessi. Nelle trascrizioni dei *files* audio, l'elisione nelle parole funzionali analizzate è indicata per mezzo di un apostrofo che sostituisce la vocale elisa. I dati del corpus sono stati sottoposti ad analisi statistica inferenziale utilizzando due test non parametrici: il Chi quadrato di Pearson ed il Test esatto di Fisher, che sono stati implementati mediante il software *SPSS 15.0* per Windows.

4.2 Le inchieste sul campo

Per le inchieste sul campo, ci si è avvalsi di 9 informatori (4 uomini e 5 donne), tutti studenti universitari di età compresa fra i 23 e 29 anni, che vivono e/o studiano a Firenze da sempre. I tre fattori testati durante le inchieste sul campo sono quelli menzionati in (3):

(3) a. La classe delle parole funzionali (determinanti vs. proclitici);
b. I tratti morfologici di genere ([ ] vs. [femminile]) e di numero
c. ([ ] vs. [plurale]) realizzati dalle vocali bersaglio di elisione;
d. La presenza/assenza di accento di parola sulla vocale iniziale di sostantivi e verbi.

Per quanto riguarda le aspettative di chi scrive in merito ai fattori in (3a) e (3b), si veda § 4.1. Per quanto concerne la presenza/assenza di accento primario sulla vocale che segue quella bersaglio di elisione, Gili-Fivela & Bertinetto (1999) hanno messo in evidenza che in italiano l'elisione vocalica in prefissazione è sfavorita in presenza di accento sulla vocale iniziale della radice. Bisol (2003) e Cabré & Prieto (2005) affermano che, nelle sequenze di due parole lessicali nel portoghese brasiliano e nel catalano, la prima vocale viene cancellata solo raramente quando la seconda vocale porta l'accento frasale. Per quanto riguarda l'elisione nelle parole funzionali nel fiorentino, ci si aspetta che l'elisione avvenga più frequentemente quando la vocale iniziale di verbi e sostantivi è atona (e non appartiene ad alcun piede metrico) e meno frequentemente se è tonica (ed appartiene ad un piede metrico).

Per verificare i contesti di applicazione dell'elisione e l'influenza dei fattori menzionati in (3), è stato utilizzato un questionario appositamente costruito. Il questionario contiene 119 sequenze di determinanti seguiti da sostantivi che iniziano per vocale e 133 sequenze di proclitici seguiti da verbi lessicali che iniziano per vocale, per un totale di 252 stimoli. I sostantivi ed i verbi lessicali selezionati possiedono le seguenti caratteristiche: a) sono quasi

esclusivamente trisillabi, per metà piani (ossia con vocale iniziale atona, appartenente ad una sillaba aperta o ad una chiusa) e per metà sdruccioli (ossia con vocale iniziale accentata, appartenente ad una sillaba aperta o ad una chiusa); b) sono da considerarsi come 'parole frequenti' con cui i parlanti dell'italiano hanno una certa familiarità;[1] c) la vocale bersaglio di elisione e la vocale iniziale di sostantivi e verbi non possiedono mai lo stesso timbro. È il caso di precisare che sono stati selezionati solo sostantivi e verbi lessicali molto/abbastanza frequenti al fine di non incorrere nella mancata applicazione dell'elisione. In effetti, precedenti studi hanno dimostrato che una bassa frequenza di occorrenza delle parole lessicali (Baroni, 2001; Gili-Fivela & Bertinetto, 1999; Bybee, 2001; Russi, 2006) e delle parole funzionali (Jurasfky *et al.,* 2001) può compromettere l'applicazione di una serie di fenomeni morfo-fonologici.

Alcuni esempi delle sequenze di parole funzionali e parole lessicali selezionate sono presentati in (4):

(4)	a.	questo amíco	questo ábito	questo attóre	questo álbero
	b.	questi amíci	questi ábiti	questi attóri	questi álberi
	b.	lo amáva	lo accénna	lo évita	lo ápplica
	c.	mi amáva	mi esclúde	mi évita	mi ámano

Le sequenze target sono state inserite all'interno di brevi frasi formate da 5-7 parole e collocate nel mezzo delle stesse; cfr. (5):

(5) a. Di sicuro questo abito è di qualità.
b. Secondo me la odiano profondamente.

I 252 stimoli sono stati somministrati agli informatori sotto forma di *slides* di una presentazione *PowerPoint*. Su ogni *slide* compariva una singola frase, distribuita su 3 righe e priva di punto finale; cfr. (6):

(6) Di sicuro
questo abito
è di qualità

Gli informatori sono stati testati singolarmente nel Dipartimento di Linguistica dell'Università di Firenze. Ciascun informatore visualizzava ogni frase contenente le forme intere delle parole funzionali per 3 secondi, poi lo stimolo spariva e dopo 1 secondo un segnale sonoro indicava all'informatore di turno di realizzare la frase visualizzata 3 secondi prima, senza poterla più visualizzare (ossia in corrispondenza di una schermata vuota). Dato uno stimolo come quello in (6), gli informatori, che erano ignari dello scopo delle inchieste, potevano realizzare le parole funzionali in forma intera (mancata elisione, cfr. 7a) o in forma elisa (elisione, cfr. 7b):

[1] I lemmi dei sostantivi e verbi lessicali selezionati possiedono una frequenza di occorrenza maggiore di 5 nelle liste di frequenza del *C-Oral-Rom* (Cresti & Moneglia, 2005) e maggiore di 2 nelle liste di frequenza del *Lip* (De Mauro *et al.*, 1993). Sono stati selezionati dei sostantivi e verbi bisillabi (invece che trisillabi), solo quando nelle liste di frequenza consultate non erano presenti dei sostantivi e verbi trisillabi dotati di adeguata frequenza di occorrenza.

(7) a. Di sicuro **questo abito** è di qualità.
b. Di sicuro **quest'abito** è di qualità.

Le 2.268 frasi realizzate dai 9 informatori (252 x 9) sono state registrate in modalità stereo su un supporto digitale (*M-Audio Micro Track 24/96 professional 2-channel mobile digital recorder*) avvalendosi di una microfono da tavolo unidirezionale (*Sony ECM-MS907*). I parametri utilizzati sono i seguenti: segnale in entrata 1/8, modalità di registrazione .wav, frequenza di campionamento 44.100 kHz. I dati elicitati sono stati importati su un computer portatile (*Dell Inspiron 1300*) e sottoposti ad analisi uditiva (numerosi ascolti iterativi) con l'ausilio del software *Audacity*. Gli stessi dati sono stati sottoposti ad analisi statistica inferenziale utilizzando il test parametrico T-test di Student per campioni appaiati, che è stato implementato con il software *SPSS 15.0* per Windows.

4.3 Motivazione dell'analisi uditiva

Un limite di questo studio è rappresentato dalla mancata analisi fonetico-acustica dei materiali discussi. Per quanto riguarda i dati elicitati sul campo, essi sono stati interamente sottoposti ad attenta analisi uditiva. Per quanto concerne i dati appartenenti al corpus, invece, l'analisi uditiva è stata effettuata solo su un campione dell'intero corpus al fine di verificare che i *files* audio fossero stati trascritti fedelmente.

Mi sembra, tuttavia, il caso di sottolineare che l'intento principale di questo studio consiste nell'investigare in modo sistematico il funzionamento dell'elisione nel fiorentino parlato e nel chiarire quali siano i fattori che determinino l'applicazione dell'elisione, piuttosto che nel chiarire le caratteristiche acustiche (valori formantici e durata) delle vocali non elise. Si è, dunque, preferito studiare un numero cospicuo di possibili contesti di elisione (4.450 nel corpus e 2.268 nelle inchieste sul campo, per un totale di 6.718 contesti) piuttosto che concentrarsi sull'implementazione delle vocali che non subiscono elisione e che possono, quindi, subire riduzione o essere realizzate come vocali 'piene' (Garrapa, 2009: 113). Si tenga anche presente che diversi studi che hanno precedentemente indagato la risoluzione delle sequenze vocaliche eterosillabiche si sono basati esclusivamente sull'analisi uditiva dei materiali esaminati (cfr., fra gli altri, Vogel *et al.*, 1983; Agostiniani,1989; Gili-Fivela & Bertinetto,1999; Cabré & Prieto, 2005; Alba, 2006; Campany, 2008).

È il caso di sottolineare che non si è ritenuto necessario sottoporre i dati elicitati all'analisi uditiva svolta da una seconda persona dal momento che la vocale bersaglio di elisione (ossia la vocale finale di determinanti e proclitici) e la vocale successiva (ossia la vocale iniziale di sostantivi e verbi lessicali) possedevano sistematicamente un timbro diverso nei contesti esaminati e, quindi, non si ponevano particolari problemi al fine di identificare l'avvenuta elisione o la mancata applicazione dell'elisione (Garrapa, 2009: 70).

5. RISULTATI

5.1 Una prima messa a fuoco: parlato spontaneo vs. parlato elicitato

I risultati emersi dall'analisi dell'elisione nel parlato spontaneo (corpus) e nel parlato elicitato (inchieste sul campo) sono molto simili. Tuttavia, come ci si aspettava, l'elisione avviene più di frequente nel parlato spontaneo; cfr. le Tabelle 7-8 e le Figure 1-2:

	Parlato *Spontaneo*		*Elicitato*	
Determinanti	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
un(o)	445/445	100	72/72	100
l(o)	812/812	100	72/72	100
una	453/468	97	133/135	98
la	701/705	99	113/117	97
quell(o)	69/69	100		
questo	110/137	80	118/126	94
quella	30/32	94		
questa	41/61	67	122/135	90
a/da/di/in/su + l(o)	642/642	100		
a/da/di/in/su + la	496/504	98		
le	13/142	9	8/135	6
quelle	2/10	20		
queste	1/5	20	28/135	21
questi	5/53	9	32/144	22
a/da/di/in/su + le	11/117	9		
Totale	3831/4202	**91**	698/1071	**65**

Tabella 7: Elisione nei determinanti nel parlato spontaneo vs. elicitato

	Parlato *Spontaneo*		*Elicitato*	
Proclitici	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
lo	25/49	51	65/180	36
la	16/21	76	78/180	43
mi	36/69	52	37/216	15
ti	14/35	40		
ci	5/28	18	90/216	42
vi	1/10	10		
li	2/14	14	16/207	8
le (acc)	0/11	0	11/189	6
le (dat)	0/11	0		
Totale	99/248	**41**	297/1188	**25**

Tabella 8: Elisione nei proclitici nel parlato spontaneo vs. elicitato

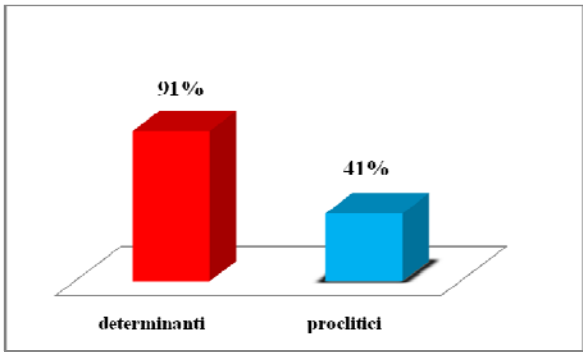


Figura 1: Elisione nel parlato spontaneo

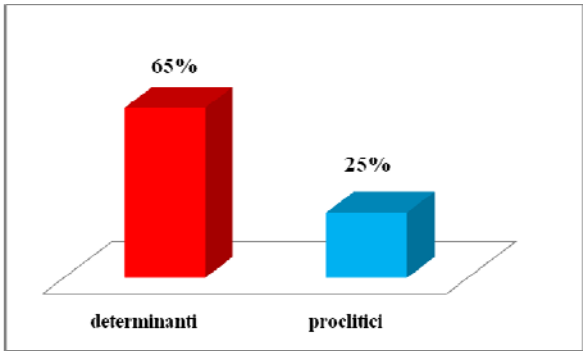


Figura 2: Elisione nel parlato elicitato

Il fatto che l'elisione venga applicata con minore frequenza nel parlato elicitato rispetto a quello spontaneo è ben noto in letteratura ed è sicuramente imputabile alla presenza scoperta degli strumenti di registrazione che incutono un certo timore (spesso involontario) negli informatori, portandoli a non elidere quelle vocali che generalmente elidono nel parlato spontaneo (Sanga, 1991; Canobbio & Telmon, 1993). Dalle Tabelle 7-8 e dalle Figure 1-2 si evince che le vocali dei determinanti vengono elise con maggiore frequenza (nel 91% nel parlato spontaneo e nel 65% dei casi nel parlato elicitato) rispetto a quelle dei proclitici (le percentuali di elisione sono del 41% nel parlato spontaneo e del 25% nel parlato elicitato).

5.2 Fattori che influenzano l'elisione

Tre dei quattro fattori testati influenzano in modo significativo l'applicazione dell'elisione nel fiorentino: la tipologia delle parole funzionali (cfr. § 5.2.1), il tratto morfologico di numero (cfr. § 5.2.2) e lo stile discorsivo (cfr. § 5.2.3).

5.2.1 La tipologia delle parole funzionali

L'influenza della tipologia delle parole funzionali sull'applicazione dell'elisione nella varietà fiorentina è stata studiata sia nel parlato spontaneo che in quello elicitato. Ne risulta che le vocali atone finali dei determinanti vengono elise con maggiore frequenza (nell'86% dei casi) rispetto a quelle dei proclitici (nel 28% dei casi); cfr. le Tabelle 9-10 e la Figura 3:

	Parlato *spontaneo* ed *elicitato*	
Determinanti	Elisione/ occ. totali	% Elisione
un(o)	517/717	100
l(o), quell(o), a/da/di/in/su + l(o)	1595/1595	100
Questo	228/263	87
Una	586/603	97
la, quella, a/da/di/in/su + la	1340/1358	99
questa	163/196	83
le, quelle, a/da/di/in/su + le	34/404	8
queste	29/140	21
questi	37/197	19
Totale	4529/5273	**86**

Tabella 9: Elisione nei determinanti nel parlato spontaneo ed elicitato

	Parlato *spontaneo* ed *elicitato*	
Proclitici	Elisione/ occ. totali	% Elisione
lo	90/209	39
la	94/201	47
mi, ti	87/320	27
ci, vi	96/254	38
li	18/221	8
le_{acc}	11/200	5
le_{dat}	0/11	0
Totale	396/1436	**28**

Tabella 10: Elisione nei proclitici nel parlato spontaneo ed elicitato

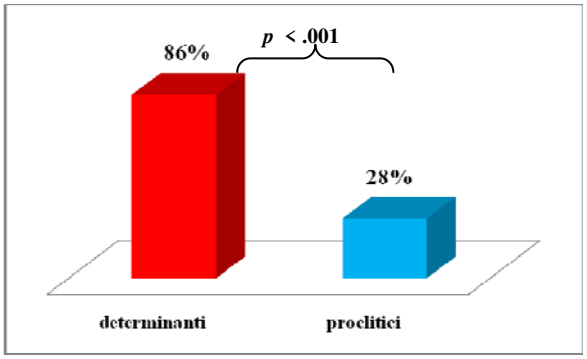


Figura 3: Elisione nel parlato spontaneo ed elicitato

La maggiore frequenza con cui le vocali atone finali dei determinanti vengono elise rispetto a quelle dei proclitici è statisticamente significativa ($p < .001$) sia nel parlato spontaneo che in quello elicitato. Ciò potrebbe essere imputabile alla ridondanza dei tratti morfologici espressi dalle vocali finali dei determinanti. In effetti, i determinanti generalmente si comportano come aggettivi rispetto ai sostantivi e le vocali finali dei determinanti

esprimono i valori di genere ([ ] vs. [femminile]) e di numero ([ ] vs. [plurale]) realizzati anche dalle vocali atone finali dei sostantivi; cfr. (8):

(8) l-a$_{f.sg.}$ oliv-a $_{f.sg.}$ quest-o $_{m.sg.}$ amic-o $_{m.sg}$

Dunque, i valori di genere e di numero realizzati dalle vocali finali dei determinanti possono considerarsi 'ridondanti'. Ne consegue che, l'elisione delle vocali finali dei determinanti non comporta la perdita di informazione morfologica. In effetti, i valori di genere e di numero realizzati dalle vocali elise possono essere ricostruiti ricorrendo al sostantivo che segue il determinante; cfr. (9):

(9) l'oliv-a $_{f.sg.}$ → l-a$_{f.sg.}$ oliv-a $_{f.sg.}$
quest' amic-o $_{m.sg}$ → quest-o $_{m.sg.}$ amic-o $_{m.sg}$

I proclitici, invece, non concordano per il genere ed il numero con il verbo che li segue. In effetti, i proclitici hanno un antecedente nel contesto linguistico o in quello discorsivo e concordano con esso per il genere ed il numero; cfr. (10):

(10) a. l-a$_{f.sg.}$ odiav-o $_{1.p.sg.}$
(*la* = individuo di sesso femminile/oggetto di genere femminile)

b. l-e$_{f.pl.}$ odiav-o $_{1.p.sg.}$
(*le* = individui di sesso femminile/oggetti di genere femminile)

Ne consegue che i valori di genere e numero espressi dalle vocali finali dei proclitici non possono essere considerati ridondanti. Dunque, l'elisione delle vocali finali atone dei proclitici comporta la perdita di informazione morfologica che non può essere immediatamente recuperata ricorrendo al verbo lessicale che segue; cfr. (11):

(11) l'odiav-o $_{1.p.sg.}$ → l-a$_{f.sg.,}$ l-e$_{f.pl.,}$ l-o$_{m.sg.,}$ l-i$_{m.pl.}$?

Dopo aver discusso l'influenza della tipologia delle parole funzionali (determinanti vs. proclitici) sull'elisione, il prossimo paragrafo esplora l'influenza dei tratti morfologici di genere e numero.

5.2.2 Il tratti morfologici di genere e numero

L'influenza esercitata dai tratti morfologici di genere ([ ] vs. [femminile]) e di numero ([ ] vs. [plurale]) sull'applicazione dell'elisione nel fiorentino è stata studiata sia nel parlato spontaneo che in quello elicitato. In linea generale, i dati congiunti mettono in evidenza che le vocali che sono sottospecificate per il tratto di numero vengono elise più frequentemente rispetto a quelle che sono esponenti morfologici del numero [plurale]. Per quanto riguarda il tratto di genere, invece, le vocali atone finali vengono elise approssimativamente con la stessa frequenza indipendentemente dal fatto che siano sottospecificate per il tratto di genere o che siano esponenti morfologici del genere [femminile].

Le vocali finali dei determinanti possono essere elise obbligatoriamente, frequentemente o raramente. Le vocali finali degli articoli *un(o)* e *l(o)* e dell'aggettivo dimostrativo

di lontananza *quell(o)* vengono sempre elise obbligatoriamente in contesto prevocalico; cfr. la Tabella 11 e gli esempi in (12). Del resto, la mancata elisione della /o/ finale di *un(o)*, *l(o)* e *quell(o)* darebbe luogo ad un *output* non grammaticale.

	Parlato *spontaneo* ed *elicitato*	
Determinanti masc. & sg.	Elisione/ occ. totali	% Elisione
un(o)	517/517	100
l(o)	884/884	100
quell(o)	69/69	100
a/da/di/in/su + l(o)	642/642	100
Totale	2112/2112	**100**

Tabella 11: Elisione nei determinanti maschili singolari nel parlato spontaneo ed elicitato

(12) l'amíco (*lo amíco) un elénco (*uno elénco) quell'attóre (*quello attóre)

Le vocali /a/ ed /o/ dei restanti determinanti singolari vengono elise molto frequentemente, ossia con una media del 96%, e presentano delle percentuali di elisione che oscillano fra l'83% per *questa* ed il 99% per *la*; cfr. la Tabella 12 e gli esempi in (13). Si noti che, in questo caso, sia le forme elise che quelle terminanti per vocale sono perfettamente grammaticali sia nel parlato che nello scritto.

	Parlato *spontaneo* ed *elicitato*	
Determinanti sg.	Elisione/ occ. totali	% Elisione
una	586/603	97
la	814/822	99
quella	30/32	94
a/da/di/in/su + la	496/504	98
questa	163/196	83
questo	228/263	87
Totale	2317/2420	**96**

Tabella 12: Elisione nei determinanti singolari nel parlato spontaneo ed elicitato

(13) un'/una idéa l'/la olíva quest'/questa offérta
quest'/questo elénco

Le vocali /e/ ed /i/ di tutti i determinanti plurali subiscono l'elisione piuttosto raramente, ossia con una media del 13% e presentano percentuali di elisione che oscillano fra l'8% ed il 21%; cfr. la Tabella 13 e gli esempi in (14). Si osservi che, sebbene sia le forme elise che quelle terminanti per vocale siano teoricamente accettabili sia nello scritto che nel parlato, le forme terminanti per vocale vengono generalmente preferite.

	Parlato *spontaneo* ed *elicitato*	
Determinanti plurali	Elisione/ occ. totali	% Elisione
le	21/277	8
quelle	2/10	20
a/da/di/in/su + le	11/117	9
queste	29/140	21
questi	37/197	19
Totale	100/741	**13**

Tabella 13: Elisione in tutti i determinanti plurali nel parlato spontaneo ed elicitato

(14) le/l'idée quelle/quell'olíve queste/quest'offérte questi/quest'amíci

Il fatto che le vocali atone finali dei determinanti singolari vengano elise quasi obbligatoriamente (ossia nel 96% dei casi, cfr. la Tabella 12), mentre le vocali finali dei determinanti plurali subiscano elisione piuttosto raramente (ossia nel 13% dei casi, cfr. la Tabella 13), risulta essere statisticamente significativo ($p < .001$) sia nel parlato spontaneo che in quello elicitato; cfr. la Figura 4:

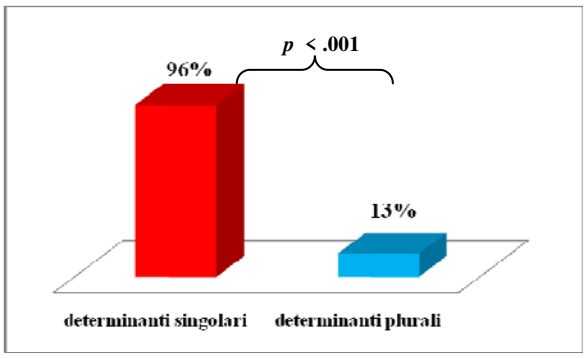


Figura 4: Elisione nei determinanti singolari vs. plurali nel parlato spontaneo ed elicitato

Sembrerebbe, dunque, che nulla impedisca la cancellazione delle vocali finali che sono sottospecificate per il tratto di numero. Al contrario, le vocali che sono esponenti morfologici del numero [plurale] tendono a rifiutare l'elisione. Per quanto riguarda il tratto di genere, /o/ ed /a/ dei determinanti singolari vengono elise con percentuali elevate (che vanno dall'83% di *questa* al 100% di *un(o), l(o)* e *quell(o)*) indipendentemente dal fatto che /o/ sia sottospecificata per il genere, mentre /a/ sia un esponente del genere [femminile]. Le vocali /i/ ed /e/ dei determinanti plurali, invece, presentano delle percentuali di elisione piuttosto basse (che vanno dall'8% di *le* al 21% di *queste*). Anche in questo caso, sembra essere del tutto irrilevante il fatto che /i/ sia sottospecificata per il genere, mentre /e/ realizzi il genere [femminile].

Per quanto riguarda i proclitici, le vocali finali dei clitici accusativi e di persona vengono elise poco frequentemente, ossia con una media inferiore al 50%. In ciò che segue, verrà discussa prima l'elisione nei clitici accusativi e poi l'elisione nei clitici di persona.

Le vocali finali dei clitici accusativi singolari sono sottospecificate per il tratto di numero (cfr. § 3) e vengono elise poco frequentemente, ossia con una media del 43%; cfr. la Tabella 14 e gli esempi in (15). Si osservi, inoltre, che la /o/ di *lo* e la /a/ di *la* subiscono elisione con una frequenza molto simile (39% vs. 47%) indipendentemente dal fatto che /o/ sia sottospecificata per il tratto di genere, mentre /a/ sia un esponente del genere [femminile].

	Parlato *spontaneo* ed *elicitato*	
Clitici acc. sing.	Elisione/ occ. totali	% Elisione
lo	94/201	47
la	90/229	39
Totale	184/430	**43**

Tabella 14: Elisione nei clitici accusativi singolari nel parlato spontaneo ed elicitato

(15) lo/l'amáva lo/l'ámano la/l'odiáva la/l'ódiano

Le vocali finali dei clitici accusativi plurali, invece, sono esponenti morfologici del numero [plurale] e vengono cancellate piuttosto raramente, ossia con una media del 5%; cfr. la Tabella 15 e gli esempi in (16). Anche in questo caso la /i/ di *li* e la /e/ di *le*$_{acc}$ subiscono elisione con una frequenza molto simile (8% vs. 5%), indipendentemente dal fatto che la /i/ di *li* sia sottospecificata per il genere, mentre la /e/ di *le*$_{acc}$ realizzi il genere [femminile].

	Parlato *spontaneo* ed *elicitato*	
Clitici acc. plur.	Elisione/ occ. totali	% Elisione
li	18/221	8
le	11/200	5
Totale	29/421	**7**

Tabella 15: Elisione nei clitici accusativi plurali nel parlato spontaneo ed elicitato

(16) li/l'offríva li/l'óffrono le/l'usáva le/l'úsano

Sebbene le vocali dei clitici accusativi subiscano l'elisione poco frequentemente, /o/ ed /a/ dei clitici accusativi singolari vengono elise con una frequenza più elevata (ossia nel 43% dei casi; cfr. la Tabella 14) rispetto ad /i/ ed /e/ dei clitici accusativi plurali (la percentuale di elisione è pari al 7%; cfr. la Tabella 15). Questo risultato è statisticamente significativo ($p < .002$) sia nel parlato spontaneo che in quello elicitato; cfr. la Figura 5:

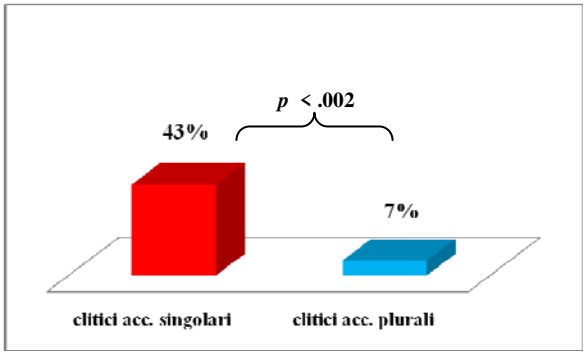


Figura 5: Elisione nei clitici accusativi singolari vs. plurali nel parlato spontaneo ed elicitato

Dunque, le vocali finali dei clitici accusativi singolari sono sottospecificate per il tratto di numero e si lasciano elidere con maggiore probabilità rispetto alle vocali finali dei clitici accusativi plurali che, invece, realizzano il numero [plurale].

Le vocali finali dei clitici di persona *mi, ti, ci, vi* e *le*, invece, non sono suffissi flessivi e, dunque, non realizzano i tratti di morfologici di genere e numero (cfr. § 3). La /i/ finale dei clitici *mi, ti, ci* e *vi* viene elisa poco frequentemente (con una media pari al 32%) ed in maniera estremamente variabile. In effetti, le percentuali di elisione di /i/ oscillano fra il 10% per *vi* ed il 40% per *ti* e *ci*; cfr. la Tabella 16 e gli esempi in (17):

	Parlato *spontaneo* ed *elicitato*	
Clitici di persona	Elisione/ occ. totali	% Elisione
mi	73/285	27
ti	14/35	40
ci	95/244	40
vi	1/10	10
Totale	183/574	**32**

Tabella 16: Elisione nei clitici di persona nel parlato spontaneo ed elicitato

(17) mi/m'odiáva mi/m'elénca
ci/c'ódiano ci/c'évita

L'elisione della /i/ finale dei clitici di persona costituisce una situazione intermedia fra l'elisione di /o/ ed /a/ dei clitici accusativi singolari e l'elisione di /i/ ed /e/ dei clitici accusativi plurali. In effetti, sebbene la /i/ di *mi, ti, ci* e *vi* non sia un affisso flessivo, essa subisce l'elisione nel 32% dei casi e, quindi, meno frequentemente rispetto ad /o/ ed /a/ di *lo* e *la*, le quali sono sottospecificate per il tratto di numero e vengono elise nel 43% dei casi (cfr. la Tabella 14). Allo stesso tempo, la /i/ di *mi, ti, ci* e *vi* viene elisa più frequentemente rispetto ad /i/ ed /e/ di *li* e le_{acc}, le quali sono esponenti morfologici del tratto di numero e vengono elise con estrema rarità, ossia nel 7% dei casi (cfr. la tabella 15). La situazione concernente l'elisione nei clitici di persona e l'elisione nei clitici accusativi singolari e plurali viene riassunta nella Figura 6.

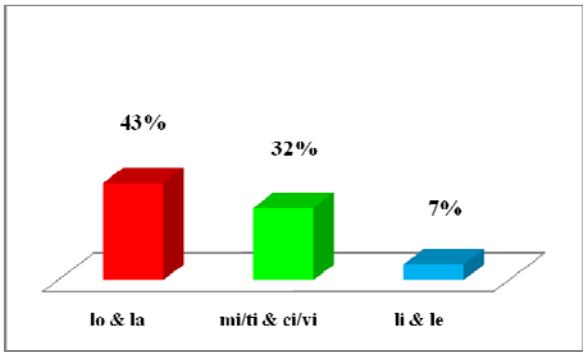


Figura 6: Elisione nei clitici di persona vs. i clitici accusativi singolari e plurali nel parlato spontaneo ed elicitato

È il caso di puntualizzare che, a differenza di quanto succede per i clitici accusativi (cfr. § 5.2.1), l'elisione della /i/ di *mi, ti, ci* e *vi* non comporta la perdita di informazione morfologica concernente i tratti di genere e numero. In primo luogo, questa /i/ non è un suffisso flessivo e, in secondo luogo, *m', t', c'* e *v'* possono essere solo le forme elise di *mi, ti, ci* e *vi*; cfr. gli esempi in (18). Al contrario, nel caso dei clitici accusativi, *l'* può essere la forma elisa di *lo, la, li e le*; cfr. gli esempi in (19).

(18) m'amáva → m' = mi_{acc}
m'elénca → m' = mi_{dat}

(19) l'amáva → l' = lo, la, li, le?
l'elénca → l' = lo, la, li, le?

In prospettiva sociolinguistica, si potrebbe pensare che la maggiore o la minore frequenza di applicazione dell'elisione nei proclitici rifletta in qualche modo la percezione che i parlanti dell'italiano contemporaneo hanno dell'applicazione dell'elisione nei contesti analizzati. A tal proposito, Agostiniani (1989: 30-32) suggerisce che la forma non elisa delle parole funzionali viene generalmente percepita come dotata di maggiore accuratezza rispetto alla sua variante elisa, che, invece, viene generalmente vista come marcata stilisticamente e, quindi, appartenente ad uno stile 'più basso'. Si potrebbe, dunque, pensare che la maggiore frequenza di applicazione dell'elisione nei proclitici accusativi singolari *lo/la* (43% dei casi) sia una conseguenza del fatto che, in questo contesto, l'elisione viene percepita come neutra dai parlanti dell'italiano contemporaneo. Inoltre, secondo le grammatiche italiane tradizionali (cfr. Dardano & Trifone, 1988; Serianni & Castelvecchi, 1988), l'elisione delle vocali /o/ ed /a/ dei proclitici *lo* e *la* è perfettamente accettabile anche nello scritto. La situazione non è la stessa per l'elisione nei proclitici di persona e nei proclitici accusativi plurali.

La /i/ finale dei clitici di persona *mi/ti* e *ci/vi* viene cancellata nel 32% dei casi, ossia meno frequentemente rispetto alle vocali dei clitici accusativi singolari (43%), ma più di frequente rispetto alle vocali dei clitici accusativi plurali (7%). Mi sembra plausibile ipotizzare che l'elisione nei proclitici di persona viene percepita come 'dotata di coloritura

regionale' (toscana, romana, umbra, ecc.) e/o associata ad uno stile piuttosto trascurato da parte dei parlanti. Secondo le grammatiche italiane (cfr. Dardano & Trifone, 1988; Serianni & Castelvecchi, 1988), l'elisione della /i/ finale dei clitici di persona è possibile (quasi) esclusivamente davanti ad un verbo che inizi per /i/. Considerando in fine l'elisione nei proclitici plurali *li/le*$_{acc}$, essa avviene piuttosto raramente (nel 7% dei casi). Sembrerebbe, dunque, che l'elisione delle vocali finali dei proclitici plurali venga percepita come arcaica da parte dei parlanti dell'italiano e/o venga associata ad uno stile estremamente trascurato. Questa percezione che i parlanti hanno potrebbe essere rafforzata dal fatto che l'applicazione dell'elisione nei clitici accusativi plurali è inusuale nella lingua scritta. Ne consegue che i parlanti potrebbero essere portati a pensare che sia preferibile evitare di elidere le vocali finali dei clitici accusativi plurali e dei clitici di persona, mentre nessuna restrizione coinvolge le vocali dei clitici accusativi singolari.

In prospettiva fonologica, il fatto che la /i/ di *mi, ti, ci* e *vi* viene elisa poco frequentemente può essere ricondotto al timbro della vocale bersaglio di elisione. La /i/ è una vocale alta anteriore e, come è ben noto in letteratura, le vocali alte possono essere realizzate sia come *glides* che come vocali piene, ossia /i/→[i] / [j] ed /u/→[u] / [w]. Le vocali medie e basse, invece, vengono realizzate come *glides* piuttosto raramente in prospettiva interlinguistica. Si deve anche notare che l'IPA dispone di un simbolo specifico per i *glides* [+ alti], ossia [j] ed [w], ma non per quelli [-alti], per i quali si può ricorrere sia al diacritico ˘ per *extrashort* che al diacritico ̯ per *non-syllabic*. La mancanza di un simbolo espressamente deputato ad indicare i *glides* [-alti] sembra rispecchiare la rarità con cui essi ricorrono in prospettiva interlinguistica. A ciò si deve aggiungere che la realizzazione delle sequenze /i.V/ sembra essere soggetta ad estrema variazione in italiano, sia quando queste sequenze ricorrono all'interno di parola che quando esse ricorrono al confine di parola (Marotta, 1987; Albano Leoni & Maturi, 2003: 44; Bertinetto & Loporcaro, 2005: 139; Canepari, 2008: 148).

Sebbene il presente studio si concentri sui casi di elisione vs. quelli di mancata elisione e non indaghi in modo specifico la maniera in cui vengono implementate le vocali non elise (cfr. § 4.3), un'osservazione è d'obbligo. L'analisi uditiva dei dati discussi (soprattutto di quelli delle inchieste sul campo) rivela che /i/ viene generalmente realizzata come un *glide* palatale quando è seguita da un sostantivo o un verbo che inizia con vocale atona (cfr. 20a), ma piuttosto come una vocale piena quando il sostantivo o il verbo seguente inizia con una vocale tonica (cfr. 20b):

(20)		mancata elisione	*gliding*	elisione
	a.	[kwest**i**#amí:tʃi]	[kwest**j**#amí:tʃi]	[kwest#amí:tʃi]
		[m**i**#amá:va]	[m**j**#amá:va]	[m#amá:va]
	b.	[kwest**i**#ábiti]		
		[m**i**#ámano]		

Quindi, la preferenza interlinguistica per le sillabe non marcate di tipo CV sembra innescare due processi paralleli, ma allo stesso tempo diversi, che hanno come target la vocale /i/, ossia il *gliding* e l'elisione. Al contrario, le vocali medie e basse vengono elise o non elise. Ne consegue che la vocale medio-alta posteriore /o/ e la vocale bassa centrale /a/ tendono a subire l'elisione con maggiore frequenza rispetto ad /i/. Come spiegare, a questo punto il fatto che la /e/ dei determinanti plurali e del clitico accusativo plurale viene elisa

solo raramente benché sia una vocale medio-alta e, come tale, non si presta ad essere realizzata come *glide*? La morfologia offre la risposta adeguata a questa domanda e, nel contempo, rivela che il timbro vocalico svolge un ruolo minore sull'elisione.

Precedentemente, si è visto che /o/ ed /a/ sono le vocali finali delle parole funzionali singolari, ossia dei determinanti *un(o)/una, l(o)/la, quell(o)/quella* e *questo/questa* e dei proclitici accusativi singolari *lo/la*. Le vocali anteriori /i/ ed /e/, invece, ricorrono in posizione finale nelle parole funzionali plurali, ossia nei determinanti *questi/queste, le/quelle* e nei proclitici accusativi plurali *li/le*. Dunque, per quanto riguarda i determinanti ed i clitici accusativi, /o/ ed /a/ sono regolarmente sottospecificate per il numero, mentre /i/ ed /e/ sono esponenti morfologici del numero [plurale]. Dal momento che la marca morfologica di numero [plurale] è alquanto resistente, la vocale centrale /a/ e la vocale posteriore /o/, che sono sottospecificate per il numero, subiscono l'elisione con maggiore probabilità rispetto ad /i/ ed /e/ che, invece, realizzano il numero [plurale].

Un aspetto interessante consiste nel fatto che /i/ subisce elisione con percentuali diverse a seconda che essa realizzi il numero [plurale], ossia nel caso del determinante plurale *questi* e del clitico accusativo plurale *li*, o che non sia un suffisso flessivo, ossia nel caso dei clitici di persona *mi, ti, ci* e *vi.* Si veda, in proposito, la Tabella 17:

Vocale /i/ finale in	Status di/i/	Elisione/ occ. totali	% Elisione
mi/ti & *ci/vi*	non è un suffisso flessivo	183/574	**32**
questi	è un suffisso flessivo	37/197	**19**
li	che realizza il numero [plurale]	18/221	**8**

Tabella 17: Elisione di /i/ a seconda del suo 'status' nel parlato spontaneo ed elicitato

Dalla Tabella 17 si evince chiaramente che la /i/ finale nei clitici di persona non è un suffisso flessivo e viene elisa più frequentemente (nel 32% dei casi) rispetto alla /i/ finale in *questi* e *li* che, invece, è un esponente morfologico del tratto di numero (la percentuale di elisione è del 19% per *questi* e dell'8% per *li*). Si può, quindi, concludere che la cancellazione delle vocali atone finali dei determinanti e dei proclitici risulta condizionata in primo luogo dalla morfologia, ossia dal tratto di numero, e solo in secondo luogo dal timbro vocalico.

Prima di terminare questa sezione, il clitico dativo femminile *le* deve essere brevemente preso in considerazione. Sebbene la /e/ finale del clitico dativo *le* non sia un suffisso flessivo (cfr. § 3), essa non viene mai elisa; cfr. la Tabella 18 e gli esempi in (21):

	Tratti morfologici				%
	Caso	*Persona*	*Numero*	*Genere*	Elisione
le	[dativo]	[ ]	[ ]	[feminine]	**0**

Tabella 18: Mancata elisione nel clitico dativo le nel parlato spontaneo

(21) le_{dat} offríva (l'_{dat}offríva) le_{dat} índica (l'índica)

La mancata elisione della /e/ finale del clitico dativo *le* non è per nulla sorprendente. In effetti, ciascun parlante dell'italiano sa perfettamente che questa /e/ non può essere elisa. Si

deve, tuttavia, notare che il clitico dativo *le* non è l'unico clitico la cui /e/ finale rifiuta l'elisione in maniera categorica pur non essendo un suffisso flessivo. In effetti, nel parlato spontaneo la /e/ finale la particella pronominale *ne* viene elisa con estrema rarità quando *ne* precede un verbo lessicale (in 1 caso su 21 nel corpus, ossia nel 5% dei casi), ad esempio in *me ne aveva già parlato → me n'aveva già parlato*. Questo ci porta a pensare che sia il clitico dativo *le* che la particella pronominale *ne* siano specificati nella loro entrata lessicale per il fatto che rifiutano l'elisione.

Dopo aver discusso l'influenza del tratto morfologico di numero sull'elisione nei determinanti e nei proclitici, la prossima sezione è dedicata allo stile discorsivo.

5.2.3 Lo stile discorsivo

Due stili discorsivi sono stati distinti nel parlato spontaneo (ossia nel corpus): il parlato informale e quello formale. Per parlato formale si intende uno stile discorsivo usato in situazioni in cui i parlanti generalmente non si conoscono (o, comunque, non sono in confidenza), che è caratterizzato da una pronuncia accurata. Per parlato informale, invece, si intende uno stile discorsivo usato in situazioni in cui i parlanti sono in confidenza e che è tendenzialmente caratterizzato da una pronuncia meno accurata rispetto a quella dello stile formale (Berruto, 1989: 13-19; Kent & Read, 2002: 227-228).

Dato che la pronuncia è generalmente più accurata nel parlato formale rispetto a quello informale, ci si aspetta che l'elisione venga messa in atto con maggiore frequenza nel parlato informale. L'analisi del corpus mette in evidenza che l'elisione risulta favorita nel parlato informale solo per un ristretto gruppo di parole funzionali, confermando così in parte l'ipotesi di chi scrive.

Chiaramente, l'elisione obbligatoria della /o/ dei determinanti maschili singolari *un(o), l(o)* e *quell(o)* non risente minimamente dello stile discorsivo; cfr. la Tabella 19.

	Parlato *spontaneo*			
	Informale		Formale	
Determinanti masc. & sg.	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
un(o)	194/194	100	251/251	100
l(o)	237/237	100	575/575	100
quell(o)	25/25	100	44/44	100
a/da/di/in/su + l(o)	181/181	100	461/461	100
Totale	637/637	**100**	1331/1331	**100**

Tabella 19: Elisione nei determinanti maschili singolari nel parlato spontaneo informale vs. formale

Anche le vocali finali dei restanti determinanti singolari vengono elise con la stessa frequenza sia nel parlato informale che in quello formale; cfr. la Tabella 20:

	Parlato *spontaneo*			
	Informale		Formale	
Determinanti singolari	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
una	260/267	97	193/201	96
la	238/240	99	463/465	99
quella	14/14	100	16/18	89
a/da/di/in/su + la	106/107	99	391/397	98
questa	17/29	59	24/32	75
questo	51/61	84	59/76	78
Totale	685/704	**97**	1146/1189	**96**

Tabella 20: Elisione nei determinanti singolari nel parlato spontaneo informale vs. formale

Le vocali finali dei determinanti plurali vengono elise poco frequentemente. Tuttavia, esse subiscono l'elisione con maggiore probabilità (nel 21% dei casi) nel parlato informale e con minore probabilità in quello formale (nel 4% dei casi); cfr. la Tabella 21. Questo risultato è statisticamente significativo ($p < .001$).

	Parlato *spontaneo*			
	Informale		Formale	
Determinanti singolari	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
le	8/34	23	5/108	4
quelle	2/8	25	0/2	0
a/da/di/in/su + le	7/44	16	4/73	5
queste	1/4	25	0/1	0
questi	5/21	24	0/32	0
Totale	23/111	**21**	9/216	**4**

Tabella 21: Elisione nei determinanti plurali nel parlato spontaneo informale vs. formale

Le vocali finali dei proclitici vengono elise poco frequentemente. Le vocali finali dei clitici accusativi singolari subiscono l'elisione con maggiore frequenza nel parlato informale (nell'85% dei casi) e con minore frequenza nel parlato formale (nel 34% dei casi). Questo risultato è statisticamente significativo ($p < .001$); cfr. la Tabella 22:

	Parlato *spontaneo*			
	Informale		Formale	
Clitici acc. singolari	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
la	13/16	81	3/5	60
lo	15/17	88	10/32	31
Totale	28/33	**85**	13/38	**34**

Tabella 22: Elisione nei clitici accusativi singolari nel parlato informale vs. formale

Le vocali finali dei clitici accusativi plurali vengono elise molto raramente. L'estrema rarità con cui i clitici accusativi plurali ricorrono nel corpus (6 occorrenze nel parlato informale e 19 nel parlato formale) non permette di determinare se lo stile discorsivo influisca sull'elisione; cfr. la Tabella 23:

	Parlato *spontaneo*			
	Informale		Formale	
Clitici acc. plurali	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
li	0/2	0	2/12	17
le	0/4	0	0/7	0
Totale	0/6	**0**	2/19	**10**

Tabella 23: Elisione nei clitici accusativi plurali nel parlato informale vs. formale

La /i/ dei clitici di persona subisce l'elisione più frequentemente nel parlato informale (nel 56% dei casi) e meno frequentemente in quello formale (nel 28% dei casi). Questo risultato è statisticamente significativo ($p < .001$); cfr. la Tabella 24:

	Parlato *spontaneo*			
	Informale		Formale	
Clitici acc. plurali	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
mi	23/37	62	13/32	41
ti	7/13	54	7/22	32
ci	2/7	28	3/21	14
vi	0/0	0	1/10	10
Totale	32/57	**56**	24/85	**28**

Tabella 24: Elisione nei clitici di persona nel parlato informale vs. formale

L'analisi dell'elisione nel parlato spontaneo ha messo in evidenza che l'elisione risente dello stile discorsivo solo in taluni contesti. Più in particolare, /i/ ed /e/ dei determinanti plurali vengono elise poco frequentemente, ma prevalentemente nel parlato informale e solo eccezionalmente in quello formale. La stessa cosa accade alle vocali /o/ ed /a/ dei clitici accusativi singolari ed alla /i/ dei clitici di persona. L'elisione obbligatoria nei determinanti maschili singolari e l'elisione frequente nei determinanti singolari, invece non viene minimamente influenzata dallo stile discorsivo.

5.3 Fattori irrilevanti: l'accento

I sostantivi ed i verbi lessicali selezionati per le inchieste sul campo sono (prevalentemente) trisillabi parossitoni o proparossitoni. La motivazione di questa scelta consiste nell'intenzione di verificare se l'elisione nei determinanti e proclitici risenta della presenza dell'accento di parola sulla vocale iniziale della parola seguente. Bisogna precisare che precedenti studi su fenomeni di cancellazione in diverse lingue romanze (cfr. Vogel *et al.*, 1983; Agostiniani, 1989; Gili-Fivela & Bertinetto, 1999; Bisol, 2003; Cabré & Prieto, 2005) e non romanze (Kager, 1997; Dehé, 2008) hanno messo in evidenza che la vocale finale di una parola lessicale o di un prefisso viene elisa prevalentemente (ma non in

maniera sistematica) se la vocale seguente è atona. Al contrario, l'elisione sarebbe quasi impossibile in presenza di una vocale seguente su cui ricada l'accento primario di parola.

Questo studio intende chiarire se la stessa situazione è valida per l'elisione nelle parole funzionali e, dunque, se le vocali finali di determinanti e proclitici vengano elise prevalentemente in presenza di un sostantivo o verbo la cui vocale iniziale sia atona. Più in particolare, si è voluto chiarire se la presenza del confine sinistro di un piede metrico fra la vocale bersaglio di elisione e la vocale seguente potesse in qualche modo inibire l'elisione. Si consideri la struttura prosodica delle sequenze di parole funzionali e lessicali presentate in (22) e (23), nelle quali i confini dei piedi metrici (trochei moraici) sono indicati dalle parentesi tonde '$(_{Ft}$)' ed i confini delle parole lessicali, che corrispondono alle parole fonologiche (*Phonological Words* = PWd) nella fonologia prosodica (Nespor & Vogel, 1986), sono indicate dalle parentesi quadre '$[_{PWd}]$':

(22) La vocale iniziale della parola lessicale è *atona*

a.	senza elisione	lo. **[$_{PWd}$** a.(**$_{Ft}$** má:).va**]**
b.	con elisione	l **[$_{PWd}$** a.(**$_{Ft}$** má:).va**]**

(23) La vocale iniziale della parola lessicale è *tonica*

a.	senza elisione	lo. **[$_{PWd}$ ($_{Ft}$** á. ma**)**.no**]**
b.	con elisione	(**$_{Ft}$** l **[$_{PWd}$** á. ma**)**.no**]**

Nell'esempio in (22) la vocale iniziale del verbo lessicale *amáva* è atona e non appartiene ad alcun piede metrico. L'applicazione o la mancata applicazione dell'elisione nell'esempio in (22) non cambia la struttura prosodica della sequenza in questione. In effetti il confine sinistro del piede metrico ed il confine sinistro della parola fonologica non sono allineati né quando l'elisione non ha luogo (cfr. 22a), né quando l'elisione ha luogo (cfr. 22b). Si noti che, se la /o/ di *lo* viene elisa, la testa consonantica *l-* viene accorpata alla sillaba iniziale del verbo diventandone l'*onset* (cfr. 22b). Dunque, l'applicazione dell'elisione non modifica la struttura prosodica in (22), dal momento che nessun segmento esterno al piede viene integrato all'interno di esso. La situazione, invece, non è esattamente la stessa nel caso in cui la vocale che segue quella bersaglio di elisione porta l'accento di parola.

Nell'esempio in (23), l'accento primario di parola ricade sulla vocale iniziale del verbo *ámano.* In questo caso la vocale iniziale del verbo appartiene ad un piede metrico e, quindi, un confine di piede separa la vocale bersaglio di elisione dalla vocale iniziale del verbo. Si osservi che, la struttura prosodica della sequenza in (23) dipende dall'applicazione o dalla mancata applicazione dell'elisione. Se l'elisione non ha luogo, il confine sinistro del piede metrico ed il confine sinistro della parola fonologica sono allineati (cfr. 23a). Se l'elisione ha luogo, invece, il confine sinistro del piede metrico e quello della parola fonologica non sono più allineati (cfr. 23b). In effetti, la testa consonantica *l-* viene introdotta all'interno del piede ed il confine del piede metrico finisce per precedere quello della parola fonologica. È bene precisare che una configurazione come quella presentata in (23a) tende a rimanere preferibilmente inalterata, mentre una configurazione come quella in (23b) viene considerata come piuttosto marcata (Peperkamp, 1997).

I dati raccolti durante le inchieste sul campo mostrano chiaramente che la presenza o l'assenza dell'accento primario di parola sulla vocale iniziale di sostantivi e verbi non ha alcun effetto sull'elisione della vocale finale dei determinanti e dei proclitici.

La /o/ degli articoli maschili singolari *un(o)* e *l(o)* viene elisa obbligatoriamente sia quando è seguita da un sostantivo che inizia per vocale atona che quando è seguita da un sostantivo che inizia per vocale tonica; cfr. la Tabella 25 e gli esempi in (24):

	Parlato *elicitato:* Elisione se la seconda vocale è Atona		Tonica	
Determinanti masc. & sg.	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
un(o)	36/36	100	36/36	100
l(o)	36/36	100	36/36	100
Totale	**72/72**	**100**	**72/72**	**100**

Tabella 25: Elisione nei determinanti maschili singolari seguiti da vocale atona vs. tonica

(24) l'amíco l'ábito

Le vocali atone finali dei restanti determinanti singolari subiscono l'elisione molto frequentemente indipendentemente dal fatto che siano seguite da una vocale atona o tonica; cfr. la Tabella 26 e gli esempi in (25):

	Parlato *elicitato:* Elisione se la seconda vocale è Atona		Tonica	
Determinanti singolari	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
una	69/72	96	44/45	98
la	71/72	99	62/63	98
questa	65/72	90	57/63	90
questo	66/72	92	52/54	96
Totale	271/288	**94**	215/225	**96**

Tabella 26: Elisione nei determinanti singolari seguiti da vocale atona vs. tonica

(25) quest'/questa olíva quest'/questa ísola

Le vocali finali di tutti i determinanti plurali vengono cancellate poco frequentemente sia quando sono seguite da una vocale atona che quando sono seguite da una vocale tonica; cfr. la Tabella 27 e gli esempi in (26):

	Parlato *elicitato:* Elisione se la seconda vocale è Atona		Tonica	
Determinanti plurali	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
le	6/72	8	2/63	3
queste	15/72	21	13/63	21
questi	15/72	21	17/72	24
Totale	36/216	**17**	32/198	**16**

Tabella 27: Elisione nei determinanti plurali seguiti da vocale atona vs. tonica

(26) quest'/queste olíve quest'/queste ísole

Si considerano ora i proclitici. Le vocali /o/ ed /a/ dei clitici accusativi singolari vengono elise con la stessa probabilità sia quando precedono un verbo lessicale che inizia con vocale tonica che quando il verbo che segue inizia con una vocale atona; cfr. la Tabella 28 e gli esempi in (27):

	Parlato *elicitato:* Elisione se la seconda vocale è Atona		Tonica	
Clitici acc. singolari	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
la	49/108	45	29/72	40
lo	33/99	33	32/81	40
Totale	82/206	**40**	61/153	**40**

Tabella 28: Elisione nei clitici accusativi singolari seguiti da vocale atona vs. tonica.

(27) lo/l'amáva lo/l'ímita

Le vocali /e/ ed /i/ dei clitici accusativi plurali subiscono l'elisione con estrema rarità. Sebbene, apparentemente l'elisione abbia luogo con maggiore probabilità davanti a verbi che iniziano con una vocale atona (la percentuale di elisione è pari al 9%) e con minore probabilità davanti a verbi che iniziano con una vocale tonica (la percentuale di elisione è pari al 4%), questo risultato non è statisticamente significativo; cfr. la Tabella 29 e gli esempi in (28):

	Parlato *elicitato:* Elisione se la seconda vocale è Atona		Tonica	
Clitici acc. singolari	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
li	10/117	8	6/90	7
le	10/99	10	1/90	1
Totale	20/216	**9**	7/180	**4**

Tabella 29: Elisione nei clitici accusativi plurali seguiti da vocale atona vs. tonica

(28) li/l'elénca lo/l'úsano

Anche la /i/ finale dei clitici di persona *mi* e *ci* sembra subire l'elisione più frequentemente davanti ad una vocale atona (nel 35% dei casi) e meno frequentemente davanti ad una vocale tonica (nel 24% dei casi). Tuttavia, anche questo risultato non è statisticamente significativo; cfr. la Tabella 30 e gli esempi in (29):

	Parlato *elicitato:* Elisione se la seconda vocale è Atona		Tonica	
Clitici di persona	Elisione/ occ. totali	% Elisione	Elisione/ occ. totali	% Elisione
mi	29/126	23	8/90	9
ci	60/126	48	30/90	33
Totale	89/252	**35**	38/180	**21**

Tabella 30: Elisione nei clitici accusativi di persona seguiti da vocale atona vs. tonica

(29) mi/m'esálta mi/m'évita

In conclusione, i dati elicitati durante le inchieste sul campo mostrano chiaramente che l'elisione nei determinanti e nei proclitici non risulta minimamente impedita dalla presenza dell'accento primario di parola sulla vocale iniziale di sostantivi e verbi. Si può, quindi, concludere che, contrariamente a quanto affermato in precedenti studi sull'elisione nelle parole lessicali ed in prefissazione (cfr. Vogel *et al.*, 1983; Agostiniani, 1989; Kager, 1997; Gili-Fivela & Bertinetto, 1999; Bisol, 2003; Cabré & Prieto, 2005; Dehé, 2008), la presenza o assenza di accento primario sulla vocale che segue la vocale bersaglio di elisione è del tutto irrilevante ai fini dell'elisione nel fiorentino.

6. DISCUSSIONE

6.1 Quanti processi di elisione nel fiorentino?

L'analisi dell'elisione nel parlato spontaneo ed elicitato ha chiarito che le vocali dei determinanti vengono elise con maggiore frequenza rispetto a quelle dei proclitici. I dati analizzati hanno portato all'individuazione di tre processi di elisione, ossia l'elisione obbligatoria in un sottogruppo di determinanti maschili singolari, l'elisione variabile nei restanti determinanti singolari e nei determinanti plurali e l'elisione poco frequente nei proclitici. Nessuno di questi tre processi di elisione risulta inibito dalla presenza dell'accento di parola sulla vocale iniziale del sostantivo o del verbo che segue la vocale bersaglio di elisione.

La vocale /o/ dei determinanti maschili singolari *un(o), l(o)* e *quell(o)* è sottospecificata per il tratto di numero e viene elisa obbligatoriamente in contesto prevocalico. Questo processo di elisione obbligatoria non risulta minimamente influenzato dallo stile discorsivo.

Le vocali della maggior parte dei determinanti vengono elise in maniera variabile. Più precisamente, /a/ ed /o/ dei determinanti singolari *una, la, quella, questa* e *questo* sono sottospecificate per il tratto di numero e vengono elise nel 96% dei casi. L'elisione quasi obbligatoria nei determinanti singolari non viene influenzata dallo stile discorsivo. Le vocali /e/ ed /i/ dei determinanti plurali *le, quelle, queste* e *questi*, invece, sono esponenti morfologici del numero [plurale] e vengono elise con una media del 13%. L'elisione poco frequente nei determinanti plurali risulta favorita nel parlato informale.

Le vocali atone finali dei proclitici subiscono elisione poco frequentemente, ma prevalentemente nel parlato informale. Le vocali /o/ ed /a/ dei clitici accusativi singolari *lo* e *la* sono sottospecificate per la categoria morfologica di numero e subiscono l'elisione nel 40% dei casi. Le vocali /i/ ed /e/ dei clitici accusativi plurali *li* e *le,* invece, realizzano il numero [plurale] e vengono elise con una media del 7%. La vocale /i/ dei clitici di persona *mi/ti* e *ci/vi* non è un suffisso flessivo e viene elisa nel 32% dei casi.

Vi è un'unica parola funzionale fra quelle analizzate che rifiuta sistematicamente l'elisione: si tratta del clitico dativo *le.*

6.2 L'elisione nei determinanti e nei proclitici come allomorfia frasale precompilata con preferenze di selezione [2]

Sulla scia di Hayes (1990), Mascaró (2007) e Bonet *et al.* (2007), in questo studio si propone di rappresentare l'elisione nei determinanti e nei proclitici nel fiorentino come un fenomeno di allomorfia frasale precompilata con delle preferenze di selezione fra gli allomorfi. Per quanto concerne il concetto di 'allomorfia frasale', seguendo Mascaró (1996, 2007), per 'allomorfia frasale' (denominata anche 'alternanza allomorfica' o 'allomorfia esterna') intendo quei contesti in cui la selezione dell'allomorfo appropriato non è condizionata internamente nel lessico, ma è determinata fonologicamente, ossia dalla forma fonologica della parola adiacente e, più precisamente, dal fatto che la parola adiacente inizi per vocale oppure per consonante. Per quanto riguarda le preferenze di selezione fra gli allomorfi, suggerisco che, per i determinanti ed i proclitici analizzati, sia le forme terminanti per vocale che quelle elise siano elencate nel lessico mentale in qualità di allomorfi e

[2] Per una discussione più approfondita si vedano i capitoli 5 e 6 in Garrapa (2009).

che le preferenze selettive fra gli allomorfi in contesto prevocalico siano indicate nelle entrate lessicali delle parole funzionali in questione.

Sebbene questa visione possa apparire poco economica in quanto induce una certa ridondanza nel lessico mentale, essa risulta perfettamente plausibile dal momento che, come rivelano le liste di frequenza del *C-Oral-Rom* (Cresti & Moneglia, 2005) e del *Lip* (De Mauro *et al.*, 1993) nonché l'intuizione dei parlanti nativi, i determinanti ed i proclitici analizzati in questo lavoro possiedono un'elevata frequenza di occorrenza nel parlato spontaneo. A tal proposito Schreuder & Baayen (1995, e lavori successivi) sostengono che le parole lascino una traccia nella memoria indipendentemente dalla loro complessità morfologica e dalla loro predicibilità e che le parole ed i gruppi di parole che vengono processati con maggiore frequenza nel lessico mentale tendono a sviluppare una propria rappresentazione lessicale.

Tornando alla rappresentazione dell'elisione nel fiorentino come un fenomeno di allomorfia frasale precompilata con preferenze selettive, si propone che le forme preferite in contesto prevocalico non siano le stesse per i determinanti singolari vs. quelli plurali, per i determinanti singolari vs. i clitici accusativi singolari e per i clitici accusativi singolari vs. quelli plurali.

6.2.1 L'elisione obbligatoria nei determinanti come allomorfia frasale precompilata categorica

L'elisione obbligatoria che trova applicazione nei determinanti maschili singolari *un(o)*, *l(o)* e *quell(o)* seguiti da sostantivi (o aggettivi) che iniziano per vocale può essere rappresentata come un caso di allomorfia frasale precompilata *categorica*. Come è ben noto, i determinanti maschili singolari hanno diverse forme superficiali che sono in distribuzione complementare. L'articolo definito maschile singolare e l'aggettivo dimostrativo di lontananza hanno tre forme superficiali: *il/lo/l'* e *quel/quello/quell'*. Le forme *il/quel* e *lo/quello* appaiono in contesto preconsonantico (cfr. 30a ed 30b), mentre le forme *l'/quell'* ricorrono in contesto prevocalico (cfr. 30c). L'articolo indefinito maschile singolare, invece, possiede solo due forme superficiali: *un* ed *uno*. La forma *un* appare sia in contesto preconsonantico che prevocalico (cfr. 31a ed 31c), mentre la forma *uno* ricorre in contesto preconsonantico (cfr. 31b):

(30)	a.	il/quel cáne	il/quel práto
	b.	lo/quello sciálle	lo/quello zío
	c.	l'/quell'amíco	l'/quell'elénco
(31)	a.	un cáne	un práto
	b.	uno sciálle	uno zío
	c.	un amíco	un elénco

Seguendo Hayes (1990) e Mascaró (2007), si suggerisce che sia le forme terminanti per vocale che quelle che terminano in consonante siano elencate nel lessico mentale come allomorfi e che gli allomorfi costituiscano dei gruppi parzialmente ordinati, in cui *il, quel* ed *un* sono gli 'allomorfi dominanti' o 'allomorfi *elsewhere*' (cfr. Kiparsky, 1973), ossia quegli allomorfi che ricorrono nel maggior numero di contesti ed hanno una certa priorità sugli altri allomorfi. I gruppi di allomorfi (parzialmente ordinati al loro interno) sono proposti in (32), in cui il simbolo '>>' significa 'ha priorità su':

(32) a. / { il >> lo, l } /
b. / { kwel >> kwel:o, kwel: } /
c. / { un >> uno } /

Dunque l''apparente' elisione obbligatoria con *l(o), quell(o)* ed *un(o)* altro non è che il risultato della selezione *categorica* degli allomorfi elisi /l/, /kwel:/ ed /un/ in contesto pre-vocalico. Sulla scia di Hayes (1990), si propone che le entrate lessicali dei determinanti maschili in questione siano quelle menzionate in (33), in cui i differenti allomorfi vi siano elencati assieme al contesto in cui ricorrono.

(33) a.

100
(proprietà sintattiche e semantiche)
/lo/ / __ [$_{[masc.] \& [sg.] N, Adj.}$ /s/+C clusters, j, t:s, d:z, ʃ:, ɲ:
/l/ / __ [$_{[masc.] \& [sg.] N, Adj.}$ V
/il/ *elsewhere*

100 = Articolo definito maschile singolare

b.

110
(proprietà sintattiche e semantiche)
/kwel:o/ / __ [$_{[masc.] \& [sg.] N, Adj.}$ /s/+C clusters, j, t:s, d:z:, ʃ:, ɲ:
/kwel:/ / __ [$_{[masc.] \& [sg.] N, Adj.}$ V
/kwel/ *elsewhere*

110 = Aggettivo dimostrativo maschile singolare di lontananza

c.

101
(proprietà sintattiche e semantiche)
/uno/ / __ [$_{[masc.] \& [sg.] N, Adj.}$ /s/+C clusters, j, t:s, d:z, ʃ:, ɲ:
/un/ *elsewhere*

101 = Articolo indefinito maschile singolare

Dal momento che la selezione dei differenti allomorfi dei determinanti in questione dipende in modo cruciale dalla forma fonologica del sostantivo (o aggettivo) che segue, si assume che i determinanti vengano istanziati fonologicamente solo dopo che i sostantivi specificati dagli stessi determinanti sono stati istanziati, come schematizzato in (34):

(34) a. Inserimento degli indici che denotano l'identità dei determinanti e dei sostantivi nella struttura sintattica

[100 $_{Det}$ 300 $_{N}$] $_{NP}$	300 = indice di *stupido*
[100 $_{Det}$ 256 $_{N}$] $_{NP}$	256 = indice di *zio*
[100 $_{Det}$ 143 $_{N}$] $_{NP}$	143 = indice di *amico*
[100 $_{Det}$ 178 $_{N}$] $_{NP}$	178 = indice di *cane*
[100 $_{Det}$ 303 $_{N}$] $_{NP}$	303 = indice di *prato*

b. *Phonological instantiation* dei sostantivi
[100 $_{Det}$ stúpido $_{N}$] $_{NP}$
[100 $_{Det}$ tsí:o $_{N}$] $_{NP}$
[100 $_{Det}$ amí:ko $_{N}$] $_{NP}$
[100 $_{Det}$ ká:ne $_{N}$] $_{NP}$
[100 $_{Det}$ prá:to $_{N}$] $_{NP}$

c. *Phonological instantiation* degli allomorfi dei determinanti
[lo $_{Det}$ stúpido $_{N}$] $_{NP}$
[lo $_{Det}$ tsí:o $_{N}$] $_{NP}$
[l $_{Det}$ amí:ko $_{N}$] $_{NP}$
[il $_{Det}$ ká:ne$_{N}$] $_{NP}$
[il $_{Det}$ prá:to $_{N}$] $_{NP}$

L'informazione contenuta nelle entrate lessicali assicura la selezione di /l/, /kwel:/ ed /un/ in contesto prevocalico, cfr. (35):

(35) l'amíco quell'amíco un amíco

Per concludere, si è proposto che l'elisione obbligatoria nei determinanti maschili singolari *l(o), quell(o)* ed *un(o)* non è il risultato di un 'vero' processo di elisione, ma consiste nella selezione *categorica* degli allomorfi elisi *l', quell'* ed *un* davanti a sostantivi (ed aggettivi) che iniziano per vocale.

6.2.2 L'elisione variabile nei determinanti come allomorfia frasale precompilata graduale

Le vocali finali dei determinanti singolari *una, la, quella, questa* e *questo* vengono elise con una frequenza pari al 96%, mentre le vocali dei corrispettivi determinanti plurali *le, quelle, queste* e *questi* subiscono elisione solo nel 13% dei casi. Si propone di rappresentare l'elisione variabile che trova applicazione nella maggior parte dei determinanti singolari ed in tutti i determinanti plurali come un processo di allomorfia frasale precompilata *graduale*, in cui le preferenze di selezione in contesto prevocalico sono elencate nelle entrate lessicali dei determinanti in questione. Tuttavia, come si vedrà, le preferenze di selezione non sono le stesse per i determinanti singolari e per quelli plurali.

Alcuni punti devono essere precisati. I determinanti in questione hanno sempre due forme superficiali, una forma che termina per vocale, ossia *una, la, quella, questa, questo, le, quelle, queste* e *questi*, ed una forma elisa, ossia *un', l', quell'* e *quest'*. I sostantivi che iniziano per consonante possono essere preceduti esclusivamente dalle forme terminanti per vocale (cfr. 36a), mentre i sostantivi che iniziano per vocale possono teoricamente essere introdotti sia dalle forme terminanti per vocale che da quelle elise (cfr. 36b):

(36) a. questa fráse questa rispósta
questi práti questi távoli
b. questa/quest'olíva questa/quest'ísola
questi/quest'amíci questi/quest'ábiti

Tuttavia, la maggiore frequenza di elisione nei determinanti singolari rispetto a quelli plurali rivela che le forme elise dei determinanti singolari vengono inserite con maggiore frequenza in contesto prevocalico rispetto alle forme elise dei determinanti plurali.

Per quanto riguarda la rappresentazione dell'elisione frequente nei determinanti singolari, seguendo Hayes (1990), Mascaró (2007) e Bonet *et al.* (2007) si suggerisce che sia le forme terminanti per vocale che quelle elise siano elencate nelle entrate lessicali dei determinanti singolari e che anche le preferenze selettive in contesto prevocalico siano codificate contestualmente nelle entrate lessicali. Più in particolare, si propone che le forme elise siano le forme preferenziali quando i determinanti singolari sono seguiti da sostantivi (o aggettivi) che iniziano per vocale, si vedano le entrate lessicali proposte in (37):[3]

(37) a.

114	
(proprietà sintattiche e semantiche)	
/kwesto/	/ __ [$_{[masc] \& [sg.] N, Adj.}$ C
/kwest/ > /kwesto/	/ __ [$_{[masc] \& [sg.] N, Adj.}$ V

114 = Aggettivo dimostrativo maschile singolare di vicinanza

b.

115	
(proprietà sintattiche e semantiche)	
/kwesta/	/ __ [$_{[fem] \& [sg.] N, Adj.}$ C
/kwest/ > /kwesta/	/ __ [$_{[fem] \& [sg.] N, Adj.}$ V

115 = Aggettivo dimostrativo femminile singolare di vicinanza

Dal momento che la selezione dei differenti allomorfi dei determinanti singolari in questione dipende in modo cruciale dalla forma fonologica del sostantivo (o aggettivo) che segue, si assume che i determinanti singolari vengano istanziati fonologicamente solo dopo che i sostantivi specificati dagli stessi determinanti sono stati istanziati (cfr. 34).

L'informazione contenuta nelle entrate lessicali rende conto del fatto che gli allomorfi elisi dei determinanti singolari rappresentano le forme preferenziali davanti a sostantivi (o aggettivi) singolari che iniziano per vocale; cfr. (38):

(38)	a.	quest'amíco	(≈ 87%)	>	questo amíco	(≈ 13%)
	b.	quest'olíva	(≈ 83%)	>	questa olíva	(≈ 17%)

Per quanto riguarda la rappresentazione dell'elisione piuttosto rara nei determinanti plurali, si assume che sia le forme terminanti per vocale che quelle elise siano contenute nelle entrate lessicali dei determinanti plurali e che le preferenze di selezione in contesto prevocalico siano elencate contestualmente nelle entrate lessicali. Al contrario di quanto proposto per i determinanti singolari (cfr. 37), si suggerisce che le forme terminanti per

[3] Al fine di evitare inutili ripetizioni, in (37) vengono indicate solo le entrate lessicali degli aggettivi dimostrativi di vicinanza. Bisogna tenere presente, però, che l'entrata lessicale di *questa* è rappresentativa anche delle entrate lessicali di *una, la* e *quella*.

vocale costituiscano le forme preferenziali quando i determinanti plurali sono seguiti da sostantivi (o aggettivi) che iniziano per vocale; cfr. le entrate lessicali proposte in (39):[4]

(39) a.

121		
(proprietà sintattiche e semantiche)		
/kwesti/	/ __ [[masc] & [pl] Nome, Agg..	C
/kwesti/ > /kwest/	/ __ [[masc] & [pl] Nome, Agg..	V

121 = Aggettivo dimostrativo maschile plurale di vicinanza

b.

122		
(proprietà sintattiche e semantiche)		
/kweste/	/ __ [[fem] & [pl] Nome, Agg..	C
/kweste/ > /kwest/	/ __ [[fem] & [pl] Nome, Agg.	V

122 = Aggettivo dimostrativo femminile plurale di vicinanza

Dal momento che la selezione dei differenti allomorfi dei determinanti plurali è condizionata dalla forma fonologica del sostantivo (o aggettivo) che segue, si assume che i determinanti plurali vengano istanziati fonologicamente solo dopo che i sostantivi specificati dagli stessi determinanti sono stati istanziati; cfr. (34).

L'informazione contenuta nelle entrate lessicali rende conto del fatto che gli allomorfi terminanti per vocale dei determinanti plurali rappresentano le forme preferenziali da inserire davanti a sostantivi plurali che iniziano per vocale; cfr. (40):

(40)	a.	questi amíci	(≈ 81%)	>	quest'amíci	(≈ 19%)
	b.	queste olíve	(≈ 79%)	>	quest'olíve	(≈ 21%)

Finora si è proposto che l'elisione variabile nei determinanti può essere rappresentata come allomorfia frasale precompilata *graduale* con delle preferenze di selezione in contesto prevocalico, le quali sono elencate nelle entrate lessicali dei determinanti. Per quanto concerne i determinanti singolari, gli allomorfi elisi costituiscono le forme preferenziali da inserire in contesto prevocalico, il che risulta nell''apparente' frequente applicazione dell'elisione. Per quanto riguarda i determinanti plurali, invece, gli allomorfi terminanti per vocale costituiscono le forme preferenziali da inserire in contesto prevocalico. Si suggerisce, inoltre, che la più frequente inserzione degli allomorfi elisi dei determinanti singolari (ma non di quelli plurali) sia condizionata dalla morfologia. In effetti, /o/ ed /a/ finali dei determinanti singolari sono sottospecificate per il tratto di numero (cfr. § 3) e, quindi, possono essere omesse più 'facilmente'. Al contrario, /i/ ed /e/ dei determinanti plurali sono esponenti morfologici del numero [plurale] (cfr. § 3) e, quindi, generalmente non vengono cancellate.

[4] Si tenga presente che l'entrata lessicale di *queste* è rappresentativa anche delle entrate lessicali di *le* e *quelle*.

6.2.3 L'elisione poco frequente nei proclitici come allomorfia frasale precompilata graduale

Le vocali finali dei clitici accusativi singolari *lo* e *la* vengono elise nel 43% dei casi, quelle dei clitici accusativi plurali *li* e *le*$_{acc}$ subiscono elisione nel 7% dei casi. La /i/ dei clitici di persona *mi, ti, ci* e *vi*, invece, viene cancellata nel 32% dei casi.

Si suggerisce che l'elisione poco frequente nei proclitici possa essere rappresentata come un processo di allomorfia frasale precompilata *graduale,* in cui le preferenze di selezione in contesto prevocalico sono elencate nelle entrate lessicali dei proclitici in questione. Come si vedrà, sebbene le preferenze di selezione in contesto prevocalico siano le stesse per tutti i clitici considerati, la morfologia è responsabile della maggiore probabilità con cui le forme elise dei clitici accusativi singolari e dei clitici di persona vengono selezionate in contesto prevocalico rispetto alle forme elise dei clitici accusativi plurali.

I proclitici esaminati hanno due forme superficiali, una forma che termina per vocale, ossia *lo, la, li, le, mi, ti, ci* e *vi*, ed una forma elisa, ossia *l', m', t', c'* e *v'*. I verbi che iniziano per consonante possono essere preceduti esclusivamente dalle forme terminanti per vocale (cfr. 41a e 41c), mentre i verbi che iniziano per vocale possono teoricamente essere introdotti sia dalle forme terminanti per vocale che da quelle elise (cfr. 41b e 41d):

(41) a. lo légge b. lo/l'amáva
c. mi salúta d. mi/m'elénca

Per quanto riguarda la rappresentazione dell'elisione poco frequente nei proclitici, seguendo Hayes (1990), Mascaró (2007) e Bonet *et al.* (2007) si suggerisce che sia le forme terminanti per vocale che quelle elise siano elencate nelle entrate lessicali dei clitici e che le forme terminanti per vocale costituiscano le forme preferenziali in contesto prevocalico. Le entrate lessicali per i proclitici analizzati sono proposte in (42):[5]

(42) a.
250		
(proprietà sintattiche e semantiche)		
/lo/	/ __	[$_{Verbo}$ C
/lo/ > /l/	/ __	[$_{Verbo}$ V

250 = Clitico accusativo maschile singolare

b.
251		
(proprietà sintattiche e semantiche)		
/la/	/ __	[$_{Verbo}$ C
/la/ > /l/	/ __	[$_{Verbo}$ V

251 = Clitico accusativo femminile singolare

[5] Si osservi che l'entrata lessicale di *mi* è rappresentativa anche delle entrate lessicali di *ti, ci* e *vi.*

c. 252
(proprietà sintattiche e semantiche)
/li/ / __ [Verbo C
/li/ > /l/ / __ [Verbo V

252 = Clitico accusativo maschile plurale

d. 253
(proprietà sintattiche e semantiche)
/le/ / __ [Verbo C
/le/ > /l/ / __ [Verbo V

253 = Clitico accusativo femminile plurale

e. 260
(proprietà sintattiche e semantiche)
/mi/ / __ [Verb C
/mi/ > /m/ / __ [Verb V

260 = Clitico di 1ª persona singolare

Dal momento che la selezione dei differenti allomorfi dei proclitici è condizionata dalla forma fonologica del verbo che segue, ossia se esso inizi per vocale o per consonante, si assume che i proclitici vengano istanziati fonologicamente solo dopo che i verbi sono stati istanziati; cfr. (34).

L'informazione contenuta nelle entrate lessicali assicura che gli allomorfi terminanti per vocale costituiscano le forme preferenziali da inserire davanti a verbi lessicali che iniziano per vocale, cfr. (43):

(43)	a.	lo amáva	(≈ 51%)	>	l'amáva	(≈ 39%)
	b.	la usáva	(≈ 53%)	>	l'usáva	(≈ 37%)
	c.	li amáva	(≈ 92%)	>	l'amáva	(≈ 8%)
	d.	le usáva	(≈ 95%)	>	l'usáva	(≈ 5%)
	e.	mi elénca	(≈ 73%)	>	m'elénca	(≈ 27%)
	f.	ci ossérva	(≈ 62%)	>	c'ossérva	(≈ 38%)

Sebbene gli allomorfi terminanti per vocale siano preferiti in contesto prevocalico, bisogna notare che gli allomorfi elisi dei clitici accusativi singolari (cfr. 43a-b) e dei clitici di persona (cfr. 43e-f) vengono selezionati con maggiore frequenza rispetto agli allomorfi elisi dei clitici accusativi plurali (cfr. 43c-d). Ancora una volta, questa situazione appare determinata dalla morfologica. In effetti, /o/ ed /a/ di *lo* e *la* sono sottospecificate per il tratto di numero ed /i/ di *mi, ti, ci* e *vi* non realizza alcun tratto morfologico, quindi possono essere omesse 'più facilmente'. Le vocali /i/ ed /e/ di *li* e le_{acc}, invece, sono esponenti morfologici del numero [plurale] e, quindi, generalmente non vengono omessi.

Per concludere, in questa sezione si è proposto di rappresentare l'elisione poco frequente nei proclitici come un processo di allomorfia frasale precompilata *graduale* in cui

gli allomorfi terminanti per vocale costituiscono le forme preferenziali da inserire in contesto prevocalico.

6.2.4 L'elisione impossibile con il clitico dativo le

La parole funzionale *le* ricopre differenti funzioni. *Le* è determinante femminile plurale, clitico accusativo femminile plurale e clitico dativo femminile singolare. La /e/ di *le* come determinante femminile plurale viene elisa nell'8% dei casi (cfr. *Tabella 13*). La /e/ di *le* come clitico accusativo plurale subisce elisione nel 5% dei casi (cfr. *Tabella 15*). In fine, la /e/ di *le* come clitico dativo femminile singolare rifiuta categoricamente l'elisione (cfr. *Tabella 18*).

Si propone che l'impossibile applicazione dell'elisione nel clitico le_{dat} sia dovuta a due fattori. Per prima cosa, a differenza di quanto si è proposto per tutti i determinanti e per i clitici accusativi e di persona (cfr. §§ 6.2.1-6.2.3) l'allomorfo eliso del clitico le_{dat} non è elencato nell'entrata lessicale del clitico in questione. Per seconda cosa, il fatto che questo clitico rifiuti categoricamente l'elisione deve essere codificato nella sua entrata lessicale. Ne consegue che l'entrata lessicale proposta per il clitico in questione potrebbe essere quella indicata in (44).

(44) $\left[\begin{array}{l} 263 \\ \text{(proprietà sintattiche e semantiche)} \\ /\text{le}/_{\text{(dativo) [- elisione]}} \end{array}\right]$

263 = Clitico dativo femminile singolare

Dopo aver considerato il fatto che il clitico dativo *le* rifiuta categoricamente l'elisione, la prossima sezione riassume le principali proposte avanzate in questo studio.

7. CONCLUSIONE

Questo lavoro ha esaminato l'applicazione dell'elisione nelle parole funzionali nella varietà di italiano parlata a Firenze. L'analisi del parlato spontaneo ed elicitato ha messo in evidenza che le vocali finali dei determinanti e dei proclitici non vengono elise con la stessa frequenza. La /o/ dei determinanti maschili singolari *un(o), l(o)* e *quell(o)* viene elisa obbligatoriamente in contesto prevocalico. Le vocali /o/ ed /a/ dei restanti determinanti singolari vengono elise molto frequentemente (96%), mentre /i/ ed /e/ dei determinanti plurali subiscono l'elisione piuttosto raramente (13%). Le vocali dei proclitici vengono elise poco frequentemente. Le vocali dei clitici *lo* e *la* subiscono elisione nel 43% dei casi, la /i/ dei clitici di persona viene elisa nel 32% dei casi e le vocali dei clitici plurali *li* e le_{acc} subiscono l'elisione solo nel 5% dei casi.

Si è proposto di rappresentare l'elisione obbligatoria, variabile e poco frequente come un fenomeno di allomorfia frasale precompilata *categorica* nel caso di *un(o), l(o)* e *quell(o),* ma *graduale* per tutte le altre parole funzionali. Per quanto riguarda l'elisione come fenomeno di allomorfia frasale precompilata *graduale*, si assume che per i determinanti e per i proclitici sia le forme terminanti per vocale che quelle elise siano elencate nel lessico mentale assieme alle preferenze selettive in contesto prevocalico. Dunque l''apparente' elisione non sarebbe il risultato di un 'vero' processo di cancellazione vocalica, ma della selezione degli allomorfi elisi in contesto prevocalico.

RINGRAZIAMENTI

Desidero ringraziare Judith Meinschaefer, Mirko Grimaldi, Janet Grijzenhout, Andrea Calabrese, Christoph Schwarze, Leonardo M. Savoia, Lori Repetti e Stefano Canalis per i loro preziosi commenti su precedenti versioni di questo lavoro. Questo studio è stato in parte finanziato dalla *DFG* attraverso il progetto di ricerca A25 'Morpho-phonological variation at word-edges: evidence from Romance' (diretto da Judith Meinschaefer) all'interno del *SFB* 471 *Variation and evolution in the Lexicon* presso l'Università di Konstanz e da una borsa di eccellenza erogata all'Autrice dal *Gleichstellungsrat* dell'Università di Konstanz. Ovviamente, eventuali errori, sviste e manchevolezze sono da attribuirsi unicamente a chi scrive.

8. BIBLIOGRAFIA

Agostiniani, L. (1989), Fenomenologia dell'elisione nel parlato in Toscana, *Rivista Italiana di Dialettologia,* 13, 3-46.

Alba, M.C. (2006), Accounting for variability in the production of Spanish vowel sequences, in *Selected Proceedings of the 9th Hispanic Linguistics Symposium* (N. Sagarra & A. Toribio, editors), Somerville, MA, Cascadilla Proceedings Project, 273-285.

Albano Leoni, F. & Maturi, P. (2003), *Manuale di fonetica*, Roma: Carocci.

Baroni, M. (2001), The representation of prefixed forms in the Italian lexicon: evidence from the distribution of intervocalic [s] and [z] in Northern Italian, *Yearbook of Morphology 1999*, 121-152.

Berruto, G. (1989), *Sociolinguistica dell'italiano contemporaneo*, Roma: La Nuova Italia.

Bertinetto, P.M. & Loporcaro, M. (2005), The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome, *Journal of the International Phonetic Association,* 35, 2, 131-151.

Bisol, L. (2003), Sandhi in Brazilian Portuguese, *Probus,* 15, 177-200.

Bonet, E., Lloret, M.R. & Mascaró, J. (2007), Allomorph selection and lexical preferences: two case studies, *Lingua,* 117, 903-927.

Bybee, J. (2001), Frequency effects on French Liaison, in *Frequency and the emergence of linguistic structure* (J. Bybee & P. Hopper, editors), Amsterdam: Benjamins, 337-359.

Cabré, T. & Prieto, P. (2005), Positional and metrical prominence effects on vowel sandhi in Catalan, in *Prosodies with special reference to Iberian languages* (S. Frota, M. Vigário & M. J. Freitas, editors), Berlin: Mouton de Gruyter, 123-157.

Campany, J. E. (2008), *Diferències fonològiques entre diversos estils de parla al caval central septentrional*. Barcelona: University of Barcelona (http://www.tdx.cat/TDX-0724108-115629).

Canepari, L. (2008), *Il MaPI. Manuale di pronuncia italiana,* Bologna: Zanichelli.

Canobbio, S. & Telmon, T. (1993), Perché un questionario?, *Atlante linguistico ed etnografico del Piemonte occidentale. Questionario I, Introduzione*, Centro stampa della regione Piemonte.

Cardinaletti, A. & Shlonsky, U. (2004), Clitic positions and restructuring in Italian, *Linguistic Inquiry,* 35, 4, 519-557.

Cresti, E. & Moneglia, M. eds. (2005), *C-ORAL-ROM. Integrated reference corpora for spoken Romance languages*, Amsterdam: Benjamins.

Dardano, M. & Trifone, P. (1988), *Grammatica italiana con nozioni di linguistica,* Bologna: Zanichelli.

De Mauro, T., Mancini, F., Vedovelli, M. & Voghera, M. (1993), *Lessico di frequenza dell'Italiano parlato*, Milano: Etaslibri.

Dehé, N. (2008), To delete or not to delete: the contexts of Icelandic final vowel deletion, *Lingua,* 118, 732-753.

Embick, D. & Noyer, R. (2005), Distributed morphology and the syntax/morphology interface, in *The Oxford Handbook of linguistic interfaces* (R. Gillian & C. Reiss, editors), Oxford: OUP, 289-324.

Farkas, D. F. (1990), Two cases of underspecification in morphology, *Linguistic Inquiry,* 21, 4, 539-550.

Garrapa, L. (2009), *Vowel elision in Florentine Italian*, Doctoral Dissertation, University of Salento & University of Konstanz.

Gili-Fivela, B. & Bertinetto, P.M. (1999), Incontri vocalici fra prefisso e radice (iato o dittongo?), *Archivio Glottologico Italiano,* 74, 2, 129-172.

Greenberg, J.H. (1966a), Some universals of grammar with particular reference to the order of meaningful elements, in *Universals of Language* (J. Greenberg, editor), Cambridge, MA: MIT Press

Greenberg, J.H. (1966b): *Language universals, with special reference to feature hierarchies*, The Hague: Mouton

Haspelmath, M. (2006), Against markedness (and what to replace it), *Journal of Linguistics,* 42, 25-70.

Hayes, B. (1990), Precompiled Phrasal Phonology, in *The Phonology-Syntax connection* (I. Sharon & D. Zec, editors), Chicago: CSLI, 85-108.

Jurafsky D., Bell, A., Gregory, M. & Raymond, W. (2001), Probabilistic relations between words: evidence from reduction in lexical production, in *Frequency and the emergence of linguistic structure* (J. Bybee & P. Hopper, editors), Amsterdam: Benjamins, 229-254.

Kager, R. (1997), Rhythmic vowel deletion in Optimality Theory, in *Derivations and constraints in phonology* (I. Roca, editor), Oxford: OUP, 463-499.

Kayne, R. (2000), Person morphemes and reflexives in Italian, French and related languages, in *Parameters and Universals* (R. Kayne, editor), Oxford: OUP, 131-162.

Kent, R. & Read, C. (2002), *The acoustic analysis of speech*, Canada: Singular Thomson Learning.

Kiparsky, P. (1973), Elsewhere in Phonology, in *A Festschrift for Morris Halle* (S. Anderson & P. Kiparsky, editors), New York: Holt, Rinehart & Winston Inc., 83-106.

Lahiri, A. & Reetz, H. (2002), Underspecified recognition, in *Laboratory Phonology* 7 (C. Gussenhoven & N. Warner, editors), Berlin: Mouton de Gruyter, 637- 685.

Marotta, G. (1987), Dittongo e iato in italiano: una difficile discriminazione, *Annali della Scuola Normale Superiore di Pisa,* 17, 3, 847-887.

Marotta, G. (1995), Apocope nel parlato di Toscana, *Studi Italiani di Linguistica Teorica e Applicata*, 24, 297-322.

Marotta, G. & Sorianello, P. (1997), Vocali contigue a confine di parola, in *Unità fonetiche e fonologiche: produzione e percezione* (P.M. Bertinetto & L. Cioni, Editors)*,* Atti delle 8^{e} Giornate di Studio del G.F.S., Pisa, 18-19 Dicembre 1997, 101-113.

Mascaró, J. (1996), External allomorphy and contractions in Romance, *Probus* 8, 181-205.

Mascaró, J. (2007), External allomorphy and lexical representation, *Linguistic Inquiry,* 38, 4, 715-735.

Nespor, M. (1990), Vowel deletion in Italian: the organization of the phonological component, *The Linguistic Review*, 7, 375-390.

Nespor, M. & Vogel, I. (1986), *Prosodic Phonology*, Dordrecht: Foris.

Newman, P. & Ratliff, M. (2001), *Linguistic fieldwork,* Cambridge: CUP.

Peperkamp, S. (1997), A representational analysis of secondary stress in Italian, *Rivista di Linguistica,* 9, 1, 189-215.

Russi, C. (2006): A usage-based account of the allomorphy of the Italian masculine definite article, *Studies in Language*, 30, 3, 575-598.

Sanga, G. (1991), I metodi della ricerca sul campo, *Rivista Italiana di Dialettologia,* 15, 165-181.

Schreuder, R. & Baayen, H. (1995), Modeling morphological processing, in *Morphological aspects of language processing* (L. B. Feldman, editor), Hillsdale: LEA, 130-153.

Serianni, L. & Castelvecchi, A. (1988), *Grammatica italiana. Italiano comune e lingua letteraria,* Torino: Utet.

Van Oostendorp, M. (1997), Style levels in conflict resolution, in *Variation, change and phonological theory* (F. Hinskens, R. van Hout & L. Wetzels, editors), Amsterdam: John Benjamins, 207-229.

Vaux, B. & Cooper, J. (1999), *Introduction to linguistic field methods*, München: Lincom Europa.

Vogel, I., Drigo, M., Moser, A. & Zannier. I. (1983), La cancellazione di vocale in italiano. *Studi di grammatica italiana*, 12, 189-230.

CONTINUUM DIAFASICO E DINAMICHE DIAGENERAZIONALI NEL BASSO E ALTO CASERTANO ORIENTALE

Edoardo Mastantuoni
Università degli Studi di Torino
emastantuoni@hotmail.com

1. PREMESSA

Si vogliono qui presentare alcuni elementi attinenti ai piani diafasico e diagenerazionale che sono emersi nella prima fase di elaborazione dei dati raccolti durante un'inchiesta dialettologica condotta in sei comuni orientali della provincia di Caserta (basso casertano: San Nicola la Strada, Castel Morrone, Ruviano; alto casertano: Baia e Latina, Piedimonte Matese, Letino), della quale si premettono qui di seguito le principali linee-guida e metodologiche.[1]

1. I sei punti sono disposti lungo un asse sud-nord, che dal circondario del capoluogo di provincia sale fino ai monti del Matese (ca. 65 km). Il materiale linguistico raccolto include l'intero continuum diafasico che unisce le varietà locali d'italiano (il substandard) ai dialetti e consente inoltre un'analisi diagenerazionale. Il lavoro finale sarà inoltre incentrato anche sugli aspetti attinenti alla diatopia, oltre che su possibili osservazioni contrastive rispetto alle *koinaí* napoletane.

2. Pur concorrendo a colmare in parte la mancanza di documentazione sui dialetti di Terra di Lavoro, l'inchiesta mira a descrivere, con una 'fotografia' delle dinamiche linguistiche in atto, gli usi linguistici di due gruppi sociali: gli anziani e i giovani.

3. Una peculiarità del campione consiste nel fatto che esso è composto quasi esclusivamente da donne, a differenza di quanto è accaduto per grandi opere geolinguistiche come l'AIS e l'ALI, nelle quali era in stragrande maggioranza maschile. È dimostrato che spesso la variazione diasessuale ha un'influenza minima rispetto a quella diagenerazionale e diastratica (oltre che alle reti sociali che ciascun parlante ha stabilito), ma la scelta di un campione femminile si è rivelata preferibile per motivi di natura pratica: le parlanti di sesso femminile si sono dimostrate più disponibili ad essere intervistate, oltre che generalmente più spigliate e loquaci.

4. Per poter essere facilmente accettato nelle comunità, e nel tentativo di stabilire un rapporto con i soggetti da intervistare quanto più possibile cordiale, rilassato e amichevole, mi sono avvalso della collaborazione di alcuni *insider*. Grazie a questi ultimi, individuandoli tra i loro parenti, amici e conoscenti, ho scelto cinque informatrici e un informatore per ciascuna delle sei località (per un totale di trentasei informatori: trenta donne e sei uomini), come illustrato nella tabella 1:

[1] Il presente lavoro costituisce l'oggetto della mia Tesi di Dottorato presso l'Università di Torino (Dottorato in Scienze del Linguaggio e della Comunicazione; indirizzo in Dialettologia italiana, Geografia linguistica e Sociolinguistica).

	≤ 40	≥ 60
F	2	3
M	1	-

Tabella 1: Costituzione numerica dei due gruppi sociali
di giovani (≤ 40) e anziani (≥ 60)

5. Le interviste sono state libere, sono state richieste adducendo un pretesto sociologico, hanno avuto una durata media di circa venti minuti e sono state registrate con dispositivo digitale visibile agli informatori.

6. Il piano descrittivo prescelto privilegia il livello fonetico: vocalismo tonico, atono, consonantismo. L'analisi del livello fonetico viene svolta con metodologia uditiva e descrittiva, ma non si escludono, in un'altra sede, possibili affondi di analisi anche strumentale su singoli aspetti di particolare complessità. Il lavoro finale considererà anche fonosintassi e prosodia, oltre agli aspetti morfologici e sintattici più rilevanti.

2. USO DEI CODICI

Pur essendo costituito solo su base diagenerazionale, il campione di fatto contiene un elemento diastratico relativo alla scolarizzazione, la quale è ovviamente molto più alta tra i parlanti sotto i quaranta, rispetto a quelli sopra i sessanta; nel panorama italiano ciò implica, per i primi, un grado proporzionalmente più elevato d'italianizzazione.

I primi risultati dell'inchiesta fanno pensare a una diacronia apparente che, più che il sistema della varietà dialettale in se stessa, coinvolge le sue relazioni con quella sovrapposta nel processo di costruzione del testo conversazionale.

La variazione dei codici nelle trentasei interviste appare piuttosto ampia: l'uso degli anziani risulta pressoché polarizzato da un lato verso il dialetto locale e conservativo (parlanti monolingui, anche se naturalmente 'contaminati' qua e là dall'italiano, a livello morfo-sintattico e lessico-semantico) e dall'altro – in misura assai minore – verso forme attenuate e arcaiche di substandard, mentre i giovani rivelano una situazione linguisticamente più composita che può essere schematizzata con tre macrocategorie:

- prevalenza di substandard;
- prevalenza di varietà mistilingui con largo uso di *code-mixing*;
- prevalenza di dialetto alternato a substandard (*code-switching*).

Per quest'ultimo tipo si può parlare di una tendenza verso la diglossia *tout court*, ossia un uso semi-esclusivo del dialetto in famiglia e tra amici (quelli presenti durante l'intervista) e dell'italiano con l'intervistatore (dove, per italiano, s'intenda sempre il substandard, con occasionale ricorso al dialetto a scopo fàtico, espressivo-enfatico o metaforico).

Nella figura d'intervistatore ho cercato di assumere una posizione diafasicamente 'neutra' avvalendomi di un substandard misto a dialetto (facendo invece ricorso a un italiano più vicino allo standard per i discorsi di natura più tecnica): ciononostante, il comportamento linguistico di alcuni soggetti giovani si è orientato ugualmente verso il polo dell'italiano, dimostrando così un condizionamento diglottico evidentemente più forte di quello adattivo o accomodativo. Ciò riguarda anche alcuni soggetti con grado d'istruzione

alto (laureandi) che hanno preferito rispondere alle domande e conversare informalmente con l'intervistatore usando l'italiano, destinando le occasionali porzioni di dialetto a una funzione meramente espressiva, enfatica o metaforica. Il dialetto compare poi nelle citazioni e nei discorsi riportati.

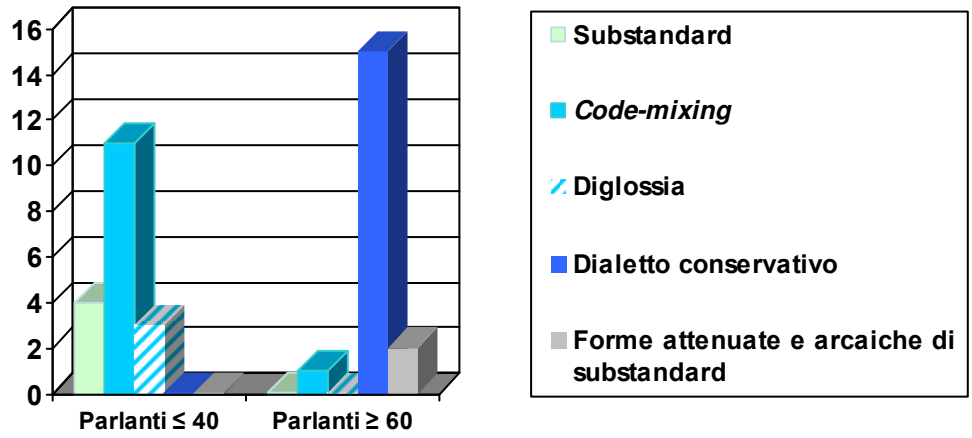


Figura 1: Rappresentazione grafica dell'uso dei codici
nelle trentasei interviste raccolte

Qui di seguito riporto alcuni esempi relativi alle osservazioni appena svolte, con i rispettivi numeri di riferimento.

Parlanti con età maggiore o uguale a 60:

- Dialetto locale conservativo: Edda, 75 anni, panettiera di Ruviano; ha seguito il primo anno di scuole elementari (file audio n. 1).
- Forme attenuate e arcaiche di substandard: Luisella, 90 anni, bracciante di Piedimonte Matese; ha conseguito la licenza elementare (file audio nn. 2a, 2b).

Parlanti con età minore o uguale a 40:

- Substandard: Marilina, 25, di Ruviano, iscritta al terzo anno del corso di laurea in Beni Culturali (file audio n. 3).
- Varietà mistilingui con largo uso di *code-mixing*: Antonietta, 25, di Ruviano, iscritta al quarto anno del corso di laurea V.O. in Psicologia (file audio n. 4).
- Dialetto e substandard (sistema diglottico: dialetto con familiari e amici; substandard con innesti dialettali con l'intervistatore): Gennaro, 24, di S. Nicola la Strada, iscritto al terzo anno d'Ingegneria (file audio n. 5. Il brano dura trenta secondi: nei primi quindici l'informatore si rivolge al cugino, negli ultimi quindici all'intervistatore).

3. EMERSIONE DI TRATTI BASILETTALI NEL LIVELLO DIAFASICO ALTO

3.1. Piano morfologico: terminazione in -a per la prima persona singolare dell'imperfetto. Indicatore o marcatore?

Nelle due sole interviste in cui due parlanti ultrasessantenni (Luisella, 90 anni, di Piedimonte Matese, e Anastasia, 79, di Baia e Latina) hanno preferito adoperare un codice substandard, sono emerse sistematicamente forme di prima persona singolare in *-a*, talvolta addirittura con posposizione del pronome personale *io* come disambiguante. Rimane da chiarire se ci si trovi davvero di fronte all'emersione di tratti basilettali nel livello diafasico alto, poiché le forme di prima persona in *-a* sono presenti anche nei dialetti dell'alto casertano (esempio: Giulia, 67 anni, di Letino: *pən'tsava*, *'ʃteva*), o se, data l'assenza di altri elementi del dialetto-base nelle formulazioni delle due parlanti, non si tratti piuttosto di una collimazione dialetto-italiano (arcaico/scolastico) che avrebbe favorito il mantenimento del tratto.

- Luisella: *sen'teva*, *io era 'sola*, *sen'tiva*, *po'tea* (file audio nn. 6-8).
- Anastasia: *M'ha aiutato tanto per questa disgrazia che aveva avuta: non veniva più a messa, io*; *io lɔ sokkor'reva*, *io an'dava*, *io* […] *por'tava* (file audio nn. 9, 10).

3.2. Piano prosodico: tratti di tipo 'laziale'

Si possono ravvisare nell'ultima informatrice citata (Anastasia, 79, di Baia e Latina) elementi intonativi analoghi a quelli del Basso Lazio (dialettologicamente e storicamente, com'è noto, parte integrante della Campania).[2] È interessante osservare come questo *pattern* intonativo si ritrovi spesso in contesti dialogici in cui compare l'avverbio *comunque* (file audio nn. 11-13a). Intonazioni simili si sono osservate a Letino per le frasi sospensive, come negli esempi appresso.

- Laura, 35: *ˌpɛntso di 'si / se ttʃ 'ɛ la 'bbwona volon'ta di 'tutti:... si po'trebbe spe'ra* (file audio n. 13b).
- Emilio (informatore *extra*), 29: *ˌin un mo'mɛnṭo deli'kaṭo ḳome 'ḳwesto...* (file audio n. 13c).

4. LA VARIAZIONE DIAGENERAZIONALE

4.1. Fricativizzazione di /tʃ/

La variabile legata all'età risulta correlata ad alcuni fenomeni riscontrati durante l'analisi dei dati, come, per esempio, la resa fricativa di /tʃ/ in [ʃ] in posizione intervocalica all'interno di parola e in fonosintassi; essa è presente in tutti giovani intervistati, ma non nei basiletti di molti anziani.

Questo tipo di variazione diagenerazionale può essere considerata il sintomo di uno sviluppo diacronico ormai in via di completamento, che comporta la prognosi di una probabile imminente perdita della variante affricata. Si tratta dunque di un fenomeno che, in un'ottica più attenta alla micro- che non alla macro-diacronia, può essere ormai considerato un elemento d'arcaicità.

[2] Vedi, per es., Canepari 1999: 433.

Il tratto converge verso un substandard genericamente centro-meridionale e, contemporaneamente, verso il dialetto parlato oggi a Napoli, che a sua volta ha attraversato lo stesso processo con una o due generazioni d'anticipo.

- Immacolata, 69, Castel Morrone: *faˈtʃevənə*, *kuˈtʃinə*, *faˈtʃevəmə*; *tʃə piˈatʃ*, *ritʃ* (file audio nn.14-18).[3]
- Gimmi, 15, Castel Morrone: *ˈfaʃilə*, *ˈdiʃə*, *viʃˈinə*, *iɱˈveʃe*, *ˈpjaʃə* (file audio nn. 19-22).

4.2. Rotacizzazione di /d/

Nei seguenti esempi è interessante notare come le due generazioni provengano dallo stesso paese (oltre che dalla stessa rete sociale). I due esiti sono stati ottenuti in un'itervista doppia.

- Carmela, 64, Ruviano: *fa ˈkauḍə* (file audio n. 23).
- Antonietta, 25, Ruviano: *fa ˈkaurə* (file audio n. 24).

4.3. Il -ne paragogico: oscillazione e lessicalizzazione

Nei parlanti più anziani, il *-ne* paragogico è produttivo – e lo si evince dal fatto che si attiva anche in parole colte come *opportunità* – tanto nell'alto quanto nel basso casertano.

- Filomena, 66 anni, Castel Morrone: *meˈtanə ~ meˈta* (file audio n. 25).
- Giulia, 67, Letino*: na ʃitˈtanə, opportuniˈta(:)nə, opportuniˈtanə, pəkˈkenə: ˌkistu ˈkkanə* (file audio nn. 26-30).

Nel gruppo dei giovani il fenomeno, ormai lessicalizzato, compare una sola volta con l'oscillazione del tratto basilettale in /ˈkkanə/:

- Vincenzino, 20, Piedimonte Matese: *pə ˈkkanə ~ pə ˈkkaːːː...* (file audio n. 31).

5. IPERCORRETTISMI

5.1. Assordimento delle occlusive sonore e livellamento sorde-sonore in una parlante di Letino già migrata al Nord e al Centro.

L'informatrice Laura, trentacinquenne, ha vissuto sette anni fuori Letino: a Milano (due anni), in Abruzzo e ad Anzio. Nei file più lunghi sono presenti anche numerosi casi di sincope e apocope di sillabe atone, dovuti naturalmente al ritmo allegro dell'eloquio. Parla anche il dialetto, ma con difficoltà.

- *bamˈbino bamˈb̥ino, ˈkwanto* («quando») *ˈsemb̥ra, ˈkwinti, ˈpren̥dere, ˈpjandʒ̊ere, p̥janˈtʃeva, ˈva̤do* (file audio nn. 32-38)

Come è udibile nel file audio 38a, *ˈsɛmp̥re k̥o ˈllɔro*, tale assordimento si accompagna con una percepibile lenizione dell'occlusiva dopo nasale (a volte quasi di tipo 'ciociaro' o 'molisano';[4] v. Vincenzino, 20, Piedimonte Matese: *g̊anˈdzoni*, *k̥anˈt̥ammo*, *ˈmango ˈt̥an̥do*, *p̥ɔ ~ b̥ɔ* «po'», *b̥uli̥do*; file audio nn. 38b, 38c e 38d) che può talora interessare alcuni parlanti di quest'area (Piedimonte Matese e Letino, per l'appunto, dove il corpus raccolto

[3] Negli esempi riportati l'affricata tende a perdere la coda vocalica in */-ə/* in finale di parola.
[4] Vedi Canepari (1999: 429 e 439).

suggerisce che si possono verificare lenizioni delle occlusive più marcate rispetto al basso casertano).

- *impa'raṭo, ḳonṭatti, pjedi'monṭe anḏare, 'vaḏo, di'pɛnte* (file audio n. 39)

5.2. Ripristino incerto di vocale finale non atona

Nel seguente esempio la parlante oscilla tra una generalizzazione del femminile in *-a* e la forma etimologica per il lessema *'retə*.

- Filomena, 66, Castel Morrone: *'reta*, *'reta* (file audio nn. 40).

6. VOCALISMO TONICO: ABBASSAMENTO DI /ɛ/ TONICA IN [æ]

Il fonema abbassato che ricorre per un solo parlante sembra esser legato soprattutto a un lessema o, più probabilmente, a una particolare variabile intonazionale e stilistica.

- Vincenzino, 20, Piedimonte Matese: *'bbæll(ə)* (abbassamento completo, file audio n. 41, e abbassamento attenuato, file audio n. 42).

7. VARIAZIONE DIAFASICA: A PROPOSITO DI 'SVIZZERA' (E ALTRI TOPONIMI)

Per via dei flussi migratori che hanno interessato tutta la zona d'inchiesta, in specie l'alto casertano, il sintagma 'in Svizzera' è comparso sovente nei discorsi dei diversi informatori; ciò ha consentito di confrontarne le molteplici varianti in cui di volta in volta si è presentato, legate non solo alla dimensione diagenerazionale, ma anche fortemente a quella diafasica.

Variabile «in Svizzera» (stato in luogo, moto a luogo):

- San Nicola la Strada
- Gennaro, 24: *in 'zvittsera*[5], *in 'ʒvittsera*, *a 'ʒvittsera* (file audio nn. 43-45).
- Maurizio (*insider*), 25: *a 'ʒvittsərə:* (file audio n. 46).
- Castel Morrone
- Immacolata, 69: *a 'ʒvittsərə* (file audio n. 47).
- Antonio (*insider*), 40 ca.: *a 'ʒvittsərə* (file audio n. 48).
- Filomena, 66, e marito (informatore *extra*), 70 ca.: *a 'ʒvittsərə* (file audio n. 49).
- Marito di Filomena (informatore *extra*), 70 ca.: *in 'zvittserə* (file audio n. 50).
- Baia e Latina
- Anastasia, 78: *in iʒ'vittsera* (file audio n. 51).
- Elvira, 85: *in iʒ'vittsera*, *a 'ʒvittsera* (file audio nn. 52a, 52b).
- Maria, 86: *in iʒ'vittsərə*, *a 'ʒvittsera* (file audio nn. 53, 54).

[5] Ho preferito, in questa sede, sillabare i nessi di fricative [zv] e [ʒv] in posizione iniziale (ossia quando non preceduti da i- prostetica), sacrificando così l'unità della sillaba fonetica, ma preservando quella della parola (per es.: *in 'zvittsera*, anziché *in z'vittsera*).

Letino

- Laura, 35: *i ˈzvittsera* (file audio n. 55).
- Marito di Filomena (informatore *extra*), 70 ca.: *in iʒˈvittsera*, *in izˈvittsera* (file audio nn. 56a, 56b).
- Filomena, 68: *in iʒˈvittsera, a la ˈʒvittsera, a la ˈʒvittsera, a la ˈʒvittsərə* (file audio n. 57-60).
- Fabio, 24: *a ˈzvittsera, in ˈzvittsera, in ˈzvittsə, a ˈzvittsera* (file audio n. 61-64).

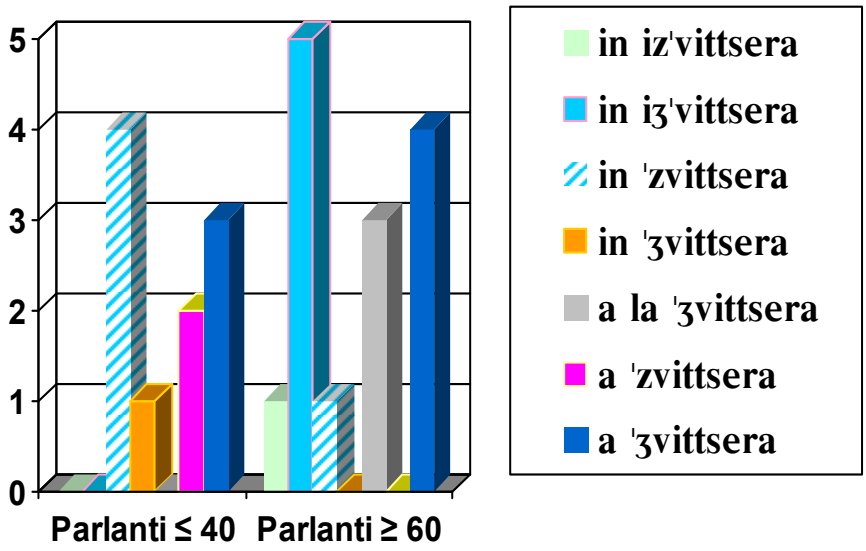


Figura 2: Rappresentazione grafica della variabile «in Svizzera» per numero di occorrenze

Dallo schema risulta chiaro che la [i-] prostetica di *liaison* è ormai legata all'uso dei parlanti anziani, i quali prediligono anche la palatalizzazione di [z] preconsonantica sia nelle varianti dialettali sia in quelle italianizzanti; a differenza dei giovani, per i quali sembra esistere una maggiore dicotomia tra la forma italiana moderna *in ˈzvittsera* e quella dialettale *a ˈʒvittsərə/-era*. Non manca ad ogni modo, per questi ultimi, qualche forma intermedia, come i due interessanti esiti *a ˈzvittsera* (prodotti durante un turno in sub-standard) con la preposizione articolata tipica del dialetto e il mantenimento del toponimo italiano. Per contro, la forma italianizzata a partire da quella dialettale, *a la ˈʒvittsera/-ərə*, è comparsa ben tre volte ma solo in una parlante appartenente alla fascia anziana.

Altre occorrenze:

«dalla Svizzera»:
- Filomena, 68, Letino: *ra la ʒvittsera* (file audio n. 65).

Risposta a: 'Di dove?':
- Maria, 78, Baia e Latina: *a ʒvittsera* (file audio n. 66).

«la Svizzera»:
- Elvira, 85, Baia e Latina: *la zvittsera*, *la ʒvittsera* (file audio nn. 67, 68).

«tutti svizzeri»:
- Fabio, 24, Letino: *zvittseri* (file audio n. 69).

(I)svizzer*	
(i)z-	10
(i)ʒ-	18

Tabella 2: Tot. occorrenze *(i)z-* vs. *(i)ʒ-*

Altri toponimi:

«a Zurigo» (forse «a Zürich»):
- Elvira, 85, Baia e Latina: *a ddzuˈrik* (file audio n. 70).

«a Zurigo […] a Schlieren»:
- Elvira, 85, Baia e Latina: *a ddzurikˈə* […] *a ʒliere* (file audio n. 71).

«Schlieren, Zurigo»:
- Elvira, 85, Baia e Latina: *ʒliere*, *ddzuˈrik(ə) (eː)* (file audio n. 72).

«a Lucerna»:
- Marito di Filomena (informatore *extra*), 70 ca., Letino: *a lluˈtʃɛrna* (file audio n. 73).

8. BIBLIOGRAFIA

AIS = K. Jaberg, J. Jud (1928-1940), Sprach und Sachatlas Italiens und der Südschweiz, Zofingen: Ringier.

ALI = *Atlante linguistico Italiano* (1995-), Roma, Istituto Poligrafico e Zecca dello Stato.

Auer, P. & Hinskens, F. (1996), The convergence and divergence of dialects in Europe: New and not so new developments in an old area, *Sociolinguistica*, 10, Tübingen: Max Niemeyer Verlag, 1-30.

Berruto, G. (1998), *Sociolinguistica dell'italiano contemporaneo*, Roma: Carocci.

Berruto, G. (2003), *Fondamenti di Sociolinguistica*, Bari: Laterza.

Canepari, L. (1999), *MaPI. Manuale di Pronuncia Italiana*, Bologna: Zanichelli.

Chambers, J.K. & Trudgill, P. (1999), *Dialectology*, Cambridge: CUP.

D'Agostino, M. (2002) (editor), *Percezione dello spazio, spazio della percezione. La variazione linguistica fra nuovi e vecchi strumenti di analisi*, Atti del Convegno internazionale, marzo 2001, Palermo (Materiali e ricerche dell'ALS, 10), Palermo: Centro studi filologici e linguistici siciliani.

De Blasi, N. (2006), *Profilo linguistico della Campania*, Bari: Laterza.

Jaberg, K. & Jud, J. (1987), *AIS. Atlante linguistico ed etnografico dell'Italia e della Svizzera Meridionale*, Milano: Unicopli.

Maturi, P. (1999), Aspetti di fonosintassi nei dialetti campani settentrionali, in *Contributi di Filologia dell'Italia Mediana*, no. XIII, 227-258.

Maturi, P. (2002), *Dialetti e substandardizzazione nel Sannio beneventano*, Francoforte: Peter Lang.

Maturi, P. & Schmid, S. (1999), Phonetically conditioned allomorphy of functional words in a dialect of Southern Italy, in *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, August 1-7, 1999, 1393-1396.

Maturi, P. & Schmid, S. (2001), Allomorfia e morfo-fonetica: riflessioni induttive su dati dialettali campani, in *Dati empirici e teorie linguistiche* (F. Albano Leoni, R. Sornicola, E. Stenta Krosbakken & C. Stromboli, editors), Atti del XXXIII Congresso Internazionale di Studi della Società di Linguistica Italiana, Napoli, 28-30 ottobre 1999, Roma: Bulzoni, 251-265.

Maturi, P. & Schmid, S. (2002), Dialettologia e fonetica acustica. Una ricerca in Campania, in *La fonetica acustica come strumento di analisi della variazione linguistica in Italia. Atti delle XII Giornate di Studio del Gruppo di Fonetica Sperimentale,* Macerata, 13-15 dicembre 2001 (A. Regnicoli, editor), Roma: il Calamo, 23-28.

Maturi, P. & Schmid, S. (2003), Sulla diffusione areale di un fenomeno di variazione morfo-fonetica nei dialetti campani, in *Actas del XXIII Congreso Internacional de Lingüística y Filología Románica*, Salamanca, 24-30 septiembre 2001 (F. Sánchez Miret, editor), Tübingen: Niemeyer, 221-233.

Milroy, L. (1989), *Observing and Analysing Natural Language: A Critical Account of Sociolinguistic Method*, Oxford: Basil Blackwell.

Radtke, E. (1997), *I dialetti della Campania*, Roma: Il Calamo.

Rohlfs, G. (1966-1969), *Grammatica storica della lingua italiana e dei suoi dialetti*. 3 voll., Torino: Einaudi.

CONFINI PROSODICI E VARIAZIONE SEGMENTALE. ANALISI ACUSTICA DELL'ALTERNANZA MONOTTONGO/DITTONGO IN ALCUNI DIALETTI DELL'ITALIA MERIDIONALE

Giovanni Abete, Adrian Simpson
Friedrich-Schiller-Universität Jena
giovanni.abete@libero.it, adrian.simpson@uni-jena.de

1. SOMMARIO

In questo contributo vengono presentati i primi risultati di un'indagine acustica, condotta su un fenomeno di alternanza sincronica tra esiti monottongali e esiti dittongali di alcune variabili vocaliche in quattro dialetti dell'Italia meridionale. L'analisi dei dati, limitata per il momento a Pozzuoli e Torre Annunziata, mette in evidenza due aspetti fondamentali di questo fenomeno, entrambi legati alla posizione delle variabili vocaliche nella struttura prosodica: l'allungamento prepausale e la tendenza delle varianti dittongali a emergere nella posizione finale di sintagma intonativo. L'alternanza monottongo/dittongo in questi dialetti rientra quindi tra i fenomeni di variazione fonetica che dipendono dalla posizione nella struttura prosodica e che forniscono al parlante indici acustici per la segmentazione della catena parlata in costituenti prosodici. Rispetto ad altre ricerche che hanno indagato il rapporto tra confini prosodici e variazione segmentale, utilizzando in genere materiale prodotto *ad hoc* in laboratorio, il presente lavoro si distingue per l'uso di parlato spontaneo relativo a varietà substandard quali i dialetti italiani.

2. INTRODUZIONE

Il presente contributo espone i primi risultati di una ricerca quadriennale che ha riguardato la variabilità del vocalismo tonico in quattro dialetti dell'Italia meridionale (Pozzuoli e Torre Annunziata in Campania, Belvedere Marittimo in Calabria, Trani in Puglia). In particolare, ci si è concentrati su un fenomeno di alternanza sincronica tra esiti monottongali e esiti dittongali, che è tra le caratteristiche più interessanti dei dialetti su menzionati. Come l'analisi evidenzierà, questa alternanza è particolarmente sensibile alla presenza di determinati confini prosodici, con le varianti dittongali che emergono in corrispondenza dei confini di ordine gerarchicamente superiore.[1] Alla questione della variazione tra esiti dittongali e esiti monottongali si aggiunge quella delle variazioni di durata, anch'esse dipendenti in larga misura dalla posizione della variabile nella struttura prosodica. Lo studio delle relazioni tra posizione prosodica, variazioni di durata e alternanza monottongo/dittongo costituisce dunque l'obiettivo primario di questo lavoro. La trattazione sarà limitata ai dati relativi a Pozzuoli e Torre Annunziata.

[1] Il rapporto tra dittongazione e posizione nella frase è stato già messo in luce da Rohlfs (1938 e 1966: § 12), il quale riporta il fenomeno come molto diffuso nei dialetti del versante adriatico, mentre sul versante tirrenico lo registra solo per i dialetti di Pozzuoli e Belvedere Marittimo. Sulla questione si veda anche Lausberg (1939: § 289) e Loporcaro (1988: 159 ss.).

Di seguito si riporta un esempio del fenomeno di alternanza in esame, attraverso alcuni brevi enunciati contenenti il lessema *rete*, tratti dalla produzione spontanea di un parlante del dialetto di Pozzuoli:[2]

1a. a̱ ˈkːosiɾ i **ɾːe̱e̱tːs** ‖ a̱ fːə̆ i **ˈʐːə̱e̱tːsə** ‖ (.)
a cucire le reti, a fare le reti (un pescatore deve imparare)

1b. t̪ʊ̯ sɑpːʰ ɪ **ɾːəetːs** ‖ ðə ˈβɑsːən̪ p̆ə vːiˈʃɣ̟ɪn̥ ‖
tu salpi le reti, ti passano (per) vicino (i motoscafi)

1c. ˈpːʊrʊ k̬wɐnd jẽn æ̃ kːəˈli i **ɾːəɪ̯tːs** ‖ (.)
pure quando andavamo a tirare le reti

2a. peˈkːʰe ˈp̆ɾɪmːə ɣʊ ˈtːʰɛndə̱ p̆jetːs ɪ **ɾetːs** zə ɣãmˈbɑ̃ː ‖ (.)
perché prima con trenta pezzi di rete si campava

2b. ɪ **ɾːetːs** ɐ mːʊˈlːˠʊ̯tːs ɛ<ɛ> (.) <ɛ> ˈnãðu ðiᵖ̆ i ɹe̱ɪtːs ‖
le reti a merluzzo è... un altro tipo di rete

2c. zæ̃n ĩ mɐgəˈdːzɪnɪ ɣɐ nu<u> tːãn i **ɾːetːs** aːˈɾɪ̲n̥t ‖
c'erano i magazzini che noi... buttavamo le reti dentro

Nei primi tre esempi le realizzazioni del lessema *rete* sono in posizione finale di sintagma intonativo, prima di pausa prosodica, e la vocale tonica presenta esiti di tipo dittongale. Negli ultimi tre esempi le realizzazioni sono in posizione interna all'enunciato, posizione che potremmo definire ora interna al sintagma fonologico, ora finale di sintagma fonologico (cfr. § 3), e presentano esiti monottongali del tipo [e]. Perché ci siano esiti dittongali la presenza di una pausa silente non sembra necessaria, come si può vedere negli esempi 1a e 1b.[3]

La situazione è comunque più complessa di quanto appaia da questi esempi. A parità di posizione prepausale gli esiti non sono sempre dittongali, ma si hanno anche esiti monottongali. Inoltre, gli stessi esiti dittongali in posizione finale sono caratterizzati da una notevole variabilità connessa alle variazioni di durata del dittongo. Il fenomeno pone quindi due problemi: da un lato descrivere correttamente l'alternanza monottongo/dittongo, dall'altro descrivere la variabilità interna agli esiti dittongali. La questione è ulteriormente complicata da un problema metodologico di fondo: distinguere tra esiti dittongali e esiti monottongali è spesso un'operazione arbitraria, se effettuata su base esclusivamente impressionistica. Gli esiti in posizione finale possono presentare talvolta traiettorie dittongali molto accentuate, ma in altri casi tali traiettorie possono essere anche solo accennate; d'altro canto, anche le realizzazioni in posizione interna non presentano andamenti formantici necessariamente piatti.[4] Pertanto, si è ritenuto necessario procedere a un'analisi acustica del fenomeno, prendendo in esame un indice numerico dell'entità della dittongazione (cfr. § 4) e studiando le variazioni di tale indice in rapporto a diverse

[2] Tra parentesi uncinate < > vengono trascritte le pause piene, dovute in genere a fenomeni di esitazione.

[3] Questa impressione è stata confermata dal confronto su base statistica delle realizzazioni vocaliche prima di pausa prosodica e prima di pausa silente, che non ha mostrato differenze significative. Per i dati si rinvia ad Abete (in preparazione).

[4] Da tempo si riconosce l'importanza del cosiddetto *Vowel Inherent Spectral Change* anche nell'identificazione di vocali di tipo monottongale (cfr. Neary & Assman, 1986).

posizioni prosodiche, evitando quindi di stabilire una distinzione impressionistica a-priori tra esiti dittongali e esiti monottongali.

Tale studio si inserisce nel filone delle ricerche che negli ultimi anni hanno indagato gli effetti della struttura prosodica non solo a livello soprasegmentale ma anche a livello segmentale (cfr. ad es. Fougeron & Keating, 1997; Keating *et al.*, 2003; Cho, 2004; Cho *et al.*, 2007). Queste ricerche hanno messo in evidenza la mole di variazione fonetica sistematica, di livello anche molto fine, che si accompagna a diverse posizioni nella struttura prosodica, attribuendo in genere a tale variazione una funzione importante nel processo di segmentazione della catena parlata in unità prosodiche di livelli diversi. Rispetto ai lavori citati, generalmente condotti su parlato di laboratorio, la presente ricerca si contraddistingue per l'uso di parlato spontaneo e di varietà substandard quali i dialetti italiani.[5] Questa scelta ha imposto una riflessione approfondita su diversi problemi metodologici, dalla modalità di elicitazione del parlato, alle tecniche di analisi acustica, ai metodi statistici per un'adeguata interpretazione dei dati. A tali aspetti si potrà in questa sede solo accennare; per una esposizione più estesa si rimanda ad Abete (in preparazione).

3. IL CORPUS

Questo studio si avvale del materiale parlato elicitato da uno stesso raccoglitore (Giovanni Abete) in una campagna di lavoro sul campo tra il 2005 e il 2007. Si tratta complessivamente di circa 28 ore di parlato di 45 informatori. Da questo corpus ampio è stato estrapolato un sotto-corpus di 24 parlanti e 20 ore di registrazioni da sottoporre all'analisi sperimentale. Come accennato, in questa sede si farà riferimento alla parte del corpus relativa a Pozzuoli e Torre Annunziata: 16 parlanti (8 per punto) per circa 12 ore di registrazioni.

Tutti i parlanti intervistati sono maschi, di età compresa tra i 30 e i 60 anni, con poche eccezioni. Sono tutti pescatori, impiegati soprattutto nel campo della piccola pesca, hanno bassa istruzione e sono spesso accomunati da condizioni di disagio economico ed emarginazione sociale.

Le interviste sono state realizzate con una versione adattata dell'intervista libera (cfr. Como, 2006). La grande maggioranza delle registrazioni è stata effettuata all'aperto, nelle baie per il rimessaggio delle barche. Qui i pescatori si intrattengono a compiere piccoli lavori di manutenzione e a rammendare le reti, trascorrendo insieme molte ore della giornata. All'intervista prendevano parte generalmente più persone. Spesso i presenti intervenivano attivamente nella conversazione, anche se non portavano personalmente il microfono.[6] L'intervistatore si rivolgeva agli intervistati in dialetto napoletano, con slittamenti verso un italiano regionale campano di livello diastraticamente basso. Il dialetto

[5] Sul rapporto tra fenomeni prosodici e andamenti formantici di un'area (quella livornese) per certi versi comparabile all'area qui in esame cfr. anche Calamai *et al.* (2003) e Marotta *et al.* (2004).

[6] La presenza all'intervista di più informatori ha attenuato il ruolo potenzialmente inibitore della presenza dell'intervistatore, favorendo dinamiche interazionali più naturali. Studiare gruppi piuttosto che individui è una delle strategie definite da Labov (1972) per attenuare i ruoli sociali di intervistatore e intervistato. Come osserva Milroy (1987: 62), "this has the effect of 'outnumbering' the interviewer and decreasing the likelihood that speakers will simply wait for questions to which they articulate".

usato dall'intervistatore, seppur diverso da quello degli intervistati, favoriva comunque l'elicitazione di parlato dialettale, in quanto veniva interpretato (rispetto all'italiano) come una varietà bassa del repertorio, e sembrava quindi autorizzare l'uso di varietà altrettanto basse. L'intervistatore partecipava attivamente alla conversazione, sia con *feedback* molto brevi (del tipo *ah*, *mh*, *eh*, *ho capito*), sia con interventi più estesi, commentando quanto detto dall'intervistato o proponendo i propri punti di vista. La conversazione veniva orientata il più possibile verso una serie di argomenti predefiniti, quali la pesca, le tecniche utilizzate, i tipi di pesci catturati, i problemi del mercato ittico, le difficoltà della vita del pescatore, eventuali avventure in mare, etc. Quando però il parlante proponeva delle digressioni rispetto a questo canovaccio di base, l'intervistatore lo assecondava, per poi riportare di nuovo il discorso sugli argomenti prestabiliti, non appena la struttura conversazionale lo avesse permesso. L'utilizzo di questo schema fisso di argomenti ha consentito la ricomparsa frequente in tutte le interviste di alcuni *items* lessicali, fornendo un base statisticamente solida per le analisi fonetiche.

Il corpus è stato segmentato ed etichettato manualmente in maniera parziale, limitando questo lavoro a una lista di parole precedentemente selezionate. La procedura è partita da un ascolto impressionistico dei materiali registrati, quindi dalla scelta di determinati *items* lessicali che presentavano variabilità degli esiti nelle vocali toniche, in particolare variabilità tra realizzazioni di tipo dittongale e realizzazioni di tipo monottongale. Sulla base degli *items* scelti sono state individuate tutte le loro realizzazioni nel corpus. Ad esempio, sono state segmentate e etichettate tutte le realizzazioni della parola *rete* nel dialetto di Pozzuoli.

La procedura di segmentazione utilizzata si basa sull'osservazione visiva dell'oscillogramma e dello spettrogramma, e sull'ascolto di porzioni di audio in corrispondenza e a cavallo dei possibili confini fonetici.[7] Una lista di criteri operativi, corredata da esempi, è stata definita preliminarmente al lavoro di segmentazione.[8]

L'etichettatura ha previsto diversi livelli: segmenti, parole, sintagmi intonativi, enunciati. Altre informazioni prosodiche sono state inserite nella trascrizione al livello segmentale. Questa procedura consente di effettuare un'analisi delle variabili fonetiche all'interno di determinati tipi lessicali, controllando la posizione che le variabili occupano in strutture prosodiche gerarchicamente più alte, come il sintagma intonativo, e fornendo informazioni sul contesto discorsivo più ampio, grazie alla trascrizione del parlato contenuto in uno o più enunciati. I dati del lavoro di etichettatura sono stati quindi trasferiti in un data-base attraverso una serie di procedure automatizzate.

In questo studio si assume una strutturazione gerarchica della prosodia, in cui i costituenti di livello superiore sono composti da costituenti di livello inferiore.[9] Di particolare importanza per l'analisi del fenomeno in esame sono i costituenti 'sintagma

[7] Il programma utilizzato per la segmentazione è Wavesurfer 1.8.5 (Sjölander & Beskow, 2006).

[8] Per le specifiche della procedura di segmentazione si rinvia ad Abete (in preparazione).

[9] Cfr. Selkirk (1984); Beckman & Pierrehumbert, J. (1986); Nespor & Vogel (1986); Selkirk (1986); Pierrehumbert & Beckman, (1988); Nespor (1993); Beckman (1996); per una rassegna Shattuck-Hufnagel & Turk (1996).

intonativo' e 'sintagma fonologico'.[10] Sulla base di questi due costituenti è possibile distinguere all'interno dei nostri materiali tra 3 diverse posizioni prosodiche (partendo dal livello più basso della gerarchia): 1) posizione interna al sintagma fonologico; 2) posizione finale di sintagma fonologico ma interna al sintagma intonativo; 3) posizione finale di sintagma intonativo. Le tre posizioni saranno indicate rispettivamente dalle sigle *SFa*, *SFb*, *SI*. In figura 1 si riporta un esempio di rappresentazione della struttura prosodica di un sintagma intonativo per un breve enunciato tratto dal corpus di Pozzuoli. Per convenienza espositiva le parole sono glossate in italiano; in basso sono riportate schematicamente le diverse posizioni prosodiche. Nello schema si evidenzia anche che le tre posizioni prosodiche possono essere raggruppate in due categorie fondamentali, opponendo da un lato la posizione finale di SI, comunemente definita 'prepausale' e dall'altro le due posizioni SFa e SFb, racchiuse in una categoria che possiamo qui definire 'interna'.

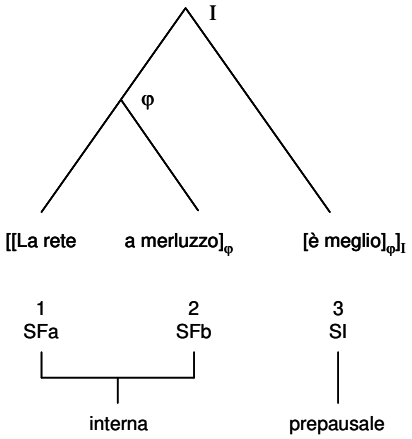


Figura 1: Albero prosodico di un sintagma intonativo tratto dal corpus di Pozzuoli. In basso si riportano le posizioni prosodiche ritenute pertinenti per la presente ricerca

Per la posizione finale di sintagma intonativo, che riveste una particolare importanza per il fenomeno in esame, sono state operate ulteriori distinzioni in base al tipo di andamento melodico di confine e alla funzione pragmatica da esso assolta. Si è quindi definita la categoria SI_Q che identifica la posizione finale di sintagma intonativo con intonazione interrogativa; la categoria SI_C che identifica la posizione finale di sintagma intonativo con la presenza di un'intonazione di 'continuazione'; la categoria SI_L che identifica la posizione finale di sintagma intonativo con la presenza di un'intonazione del tipo 'lista'. Un'etichetta SI_N è stata creata ad indicare una categoria 'cestino', che accoglie le posizioni finali di sintagma intonativo prive dei toni ascendenti caratteristici delle categorie

[10] Il sintagma intonativo viene qui definito come la sequenza di parlato inclusa tra due pause prosodiche di livello superiore (cfr. Nespor & Vogel 1986; Nespor 1993). La definizione di sintagma fonologico, invece, si rifà a quella di *intermediate phrase* fornita da Beckman & Pierrehumbert (1986): un contorno intonativo con uno o più *pitch accents*, ma privo di tono di confine finale.

SI_Q, SI_C, SI_L; si tratta quindi di una categoria piuttosto ampia che abbraccia diverse funzioni pragmatiche: può ad esempio essere caratterizzata da un tono di conclusione (soprattutto quando seguita da una pausa silente), ma può avere anche altre funzioni, come quella vocativa o quella imperativa.

È evidente che questa ulteriore suddivisione della posizione finale di sintagma intonativo avrebbe potuto comportare l'individuazione di un numero molto maggiore di sottocategorie. In questa sede però si è preferito separare solo quelle categorie che sembravano avere effetti più macroscopici sulla durata e sulla presenza e l'entità delle dittongazioni, evitando un proliferare di categorie, che avrebbe comportato un'eccessiva frammentazione dei dati e la comparsa di categorie documentate da pochissimi *tokens* e difficilmente utilizzabili nei ragionamenti statistici.

Infine, un'ulteriore categoria SI_V è stata creata per identificare quelle parole che, seppur in posizione finale di sintagma intonativo e caratterizzate da un tono di confine, sono realizzate con una notevole velocità di eloquio, senza il rallentamento tipico della posizione prepausale. Questa peculiarità determina per tale categoria una durata delle toniche in posizione finale molto più bassa della media (cfr. figg. 7-8) e un'incidenza quasi nulla dei fenomeni di dittongazione (cfr. figg. 9-10). In figura 2 si evidenziano tutte le posizioni prosodiche definite per questa ricerca. Sono stati esclusi, invece, dalle rappresentazioni grafiche e dai ragionamenti statistici i *tokens* caratterizzati da allungamenti anomali dovuti a fenomeni di esitazione[11].

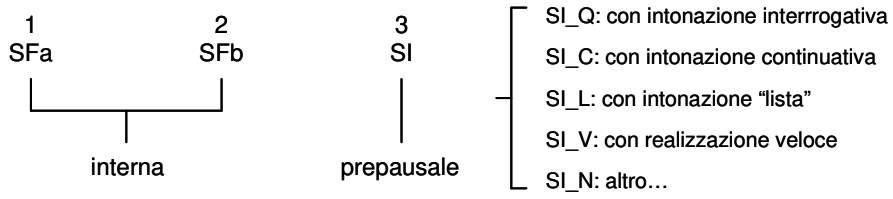


Figura 2: Posizioni nella struttura prosodica e sottocategorie di SI

Sulla base dei *tokens* vocalici etichettati sono state effettuate analisi acustiche della durata e della struttura formantica. Nella tabella 1 si riporta il numero di *tokens* per variabile vocalica e per parlante, analizzati per Pozzuoli e Torre Annunziata. Sono stati etichettati e analizzati 944 *tokens* per il dialetto di Pozzuoli e 799 per Torre Annunziata. Non tutte le variabili vocaliche e non tutti i parlanti sono rappresentati allo stesso modo. In particolare, si rileva un numero nettamente inferiore di realizzazioni di /o/ e /u/, che dipende dalla minore frequenza assoluta di queste vocali nei dialetti esaminati.

Si tenga presente che l'alternanza monottongo/dittongo non riguarda esattamente lo stesso numero di variabili nelle varietà esaminate: mentre a Torre Annunziata sono coinvolte dal fenomeno in questione tutte le vocali a eccezione della /a/, a Pozzuoli sono

[11] I dati di durata e del coefficiente di dittongazione di tali *tokens* sono riportati comunque nelle tabelle in appendice sotto l'etichetta SI_H.

coinvolte le alte e le medio-alte anteriori e posteriori.[12] In entrambi i dialetti la presenza delle varianti dittongali non è limitata né dal tipo sillabico, né dalla struttura accentuale della parola, quindi i dittonghi compaiono anche in sillaba chiusa e in parole ossitone e proparossitone.[13]

Pozzuoli							Torre Annunziata							
parlante	/i/	/e/	/ɛ/	/o/	/u/	tot.	parlante	/i/	/e/	/ɛ/	/ɔ/	/o/	/u/	tot.
PZ01	62	62	65	14	22	225	TA01	33	53	42	30	12	11	181
PZ03	13	41	39	13	12	118	TA03	34	37	18	34	9	10	142
PZ05	19	24	42	23	4	112	TA05	5	21	19	11	3	7	66
PZ07	34	6	46	7	4	97	TA07	4	8	26	37	7	1	83
PZ09	20	21	27	0	8	76	TA09	18	24	40	22	5	0	109
PZ11	14	52	51	10	15	142	TA11	4	15	30	22	3	3	77
PZ13	21	23	34	5	5	88	TA13	1	16	20	14	7	4	62
PZ15	26	9	30	6	15	86	TA15	8	17	26	24	2	2	79
tot.	209	238	334	78	85	944	tot.	107	191	221	194	48	38	799

Tabella 1: Numero di tokens analizzati per variabile vocalica e per parlante

4. CARATTERIZZAZIONE DELLA DINAMICA DITTONGALE

Mentre le durate sono state ottenute direttamente dai *files* di etichettatura, per l'analisi della traiettoria dittongale è stata realizzato uno *script* in *Snack* e *tcl/tk* per la stima automatica dei valori formantici. Il metodo di caratterizzazione della traiettoria dittongale utilizzato nella presente ricerca costituisce una evoluzione del metodo di Holbrook & Fairbancks (1962) e segue essenzialmente Simpson (1998), con qualche differenza rispetto all'algoritmo per la stima dei valori formantici. Un esempio di tale procedura è dato in figura 3.

Di ciascun segmento vocalico vengono scartati in automatico i primi 20 ms. e gli ultimi 20 ms.; ciò è necessario per ridurre l'impatto delle transizioni formantiche, che nei criteri di segmentazione utilizzati in questa ricerca vengono incluse nel segmento vocalico. La prima e l'ultima misurazione delle formanti vengono effettuate subito dopo i primi 20 ms. e subito prima degli ultimi 20 ms. I restanti punti in cui effettuare le misurazioni vengono calcolati dividendo in parti uguali la porzione tra la prima e l'ultima misurazione, facendo in modo che la lunghezza dell'unità di segmentazione sia la più vicina possibile al valore nominale di 20 ms. Quindi vengono stimati i valori delle prime tre formanti per ciascuno dei punti precedentemente individuati. Ogni segmento vocalico inferiore ai 60 ms. viene trattato tecnicamente come un monottongo e viene caratterizzato da un singolo insieme di valori di F1, F2 e F3 presi nella porzione centrale della vocale, al centro di una finestra di 20 ms.

Nell'esempio riportato in figura 3 un dittongo di 192 ms. viene rappresentato da 9 misurazioni delle formanti. Esclusi i primi e gli ultimi 20 ms., le misurazioni vengono effettuate nel punto iniziale del segmento, quindi a intervalli regolari della durata nominale di 20 ms. (la durata effettiva in questo caso è di 19 ms.).

[12] In questo dialetto l'alternanza monottongo/dittongo si va estendendo anche alla vocale anteriore medio-bassa nei pescatori al disotto dei 40 anni. Per la questione si rinvia a Abete & Simpson (in stampa).

[13] Sulla questione dei dittonghi in sillaba chiusa cfr. Abete (2006).

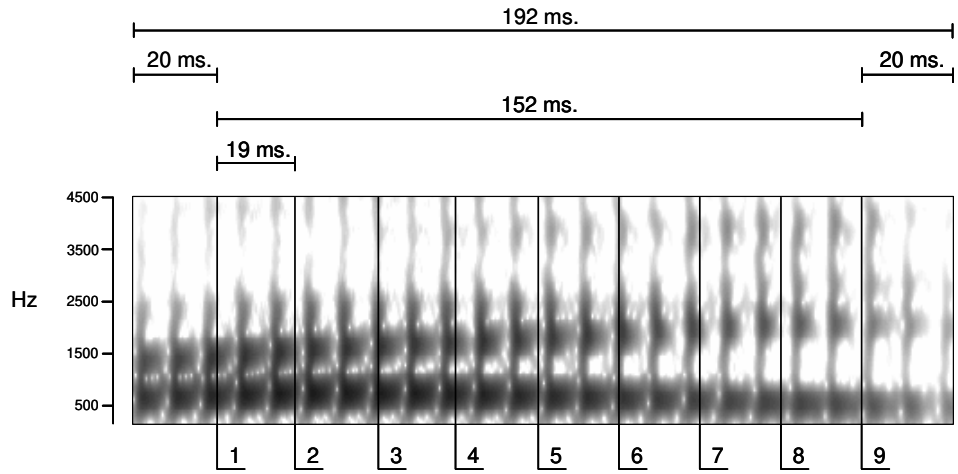


Figura 3: Esempio della procedura di analisi della traiettoria dittongale

Di seguito si riportano le specifiche tecniche per l'analisi formantica:

framelength = 0.02 s.
windowlength = 0.02
windowtype = Hamming
preemphasisfactor = 0.9
lpc-order = 12
ds_freq = 10000 Hz
nom_f1_freq = 500 Hz

Una volta effettuata la stima automatica dei valori formantici è stato necessario controllare manualmente questi valori e correggerli nel caso di errori. In questa ricerca la presenza di eventuali errori è stata individuata in maniera visiva, attraverso l'osservazione di grafici realizzati per ciascun segmento vocalico. I grafici sono stati ottenuti in automatico, grazie a uno *script* appositamente realizzato con il programma per l'analisi statistica e la rappresentazione grafica dei dati *R*.[14] Questo script realizza grafici delle traiettorie dittongali (basati sull'andamento temporale di F1 e F2), aggiungendo a ciascun grafico una serie di informazioni: codice del file, posizione prosodica, tipo di variabile, durata in ms., coefficiente di dittongazione (cfr. più avanti), la parola da cui il *token* è tratto. Il tutto viene sistemato in un unico file *.pdf, includendo dieci grafici per pagina (v. figura 4).

Se si suddividono i segmenti vocalici in classi di durata di 20 ms., i dittonghi di una stessa classe saranno caratterizzati da un eguale numero di misurazioni formantiche. È quindi possibile calcolare le medie delle misurazioni formantiche di tutti i *tokens* di un dittongo che appartengono a una stessa classe di durata, in modo da poter rappresentare graficamente le traiettorie dittongali medie per le diverse classi di durata dei *tokens* di uno

[14] R Core Development Team (2008).

stesso dittongo. Un esempio di questa modalità di rappresentazione è riportato in figura 5. Questo tipo di grafico consente di visualizzare il rapporto tra struttura del movimento dittongale e variazioni di durata. In questo caso, in particolare, si nota una notevole variabilità del timbro di attacco del dittongo in rapporto alle diverse classi di durata, con le due formanti in attacco che si avvicinano sempre più nei dittonghi di durata maggiore.[15]

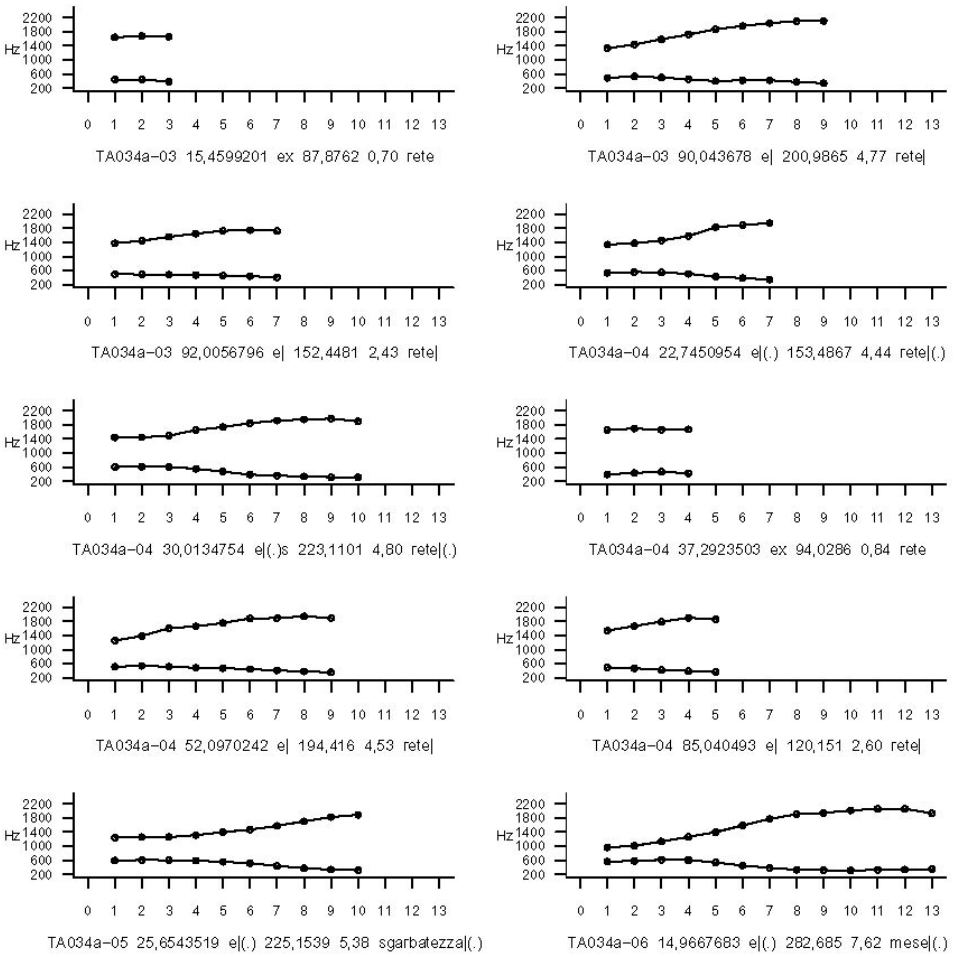


Figura 4: Grafici riassuntivi delle caratteristiche acustiche dei tokens analizzati. Esempio tratto dal corpus di Torre Annunziata e relativo a 10 realizzazioni di /e/

Oltre ad avere una rappresentazione grafica delle traiettorie formantiche, si è avuta la necessità di ottenere un indice numerico relativo all'ampiezza del movimento dittongale, così da poter confrontare l'andamento di tale indice rispetto ad altri parametri linguistici,

[15] Sulla variabilità delle traiettorie dittongali in rapporto a variazioni di durata non possiamo soffermarci in questa sede; per la questione si rinvia ad Abete (in preparazione).

come la posizione della variabile nella struttura prosodica (cfr. § 5.2), o extralinguistici, come l'età dei parlanti (cfr. Abete & Simpson, in stampa). L'indice misura la distanza euclidea nello spazio F1-F2 che intercorre tra i due timbri più distanti raggiunti dal segmento nella transizione dittongale. La base della procedura di calcolo sta nell'individuazione degli scarti tra valore massimo e valore minimo di ciascuna formante (F1 e F2) per le misurazioni effettuate sull'intero dittongo. Si ottengono così due misure separate indicative dell'ampiezza delle transizioni delle prime due formanti.

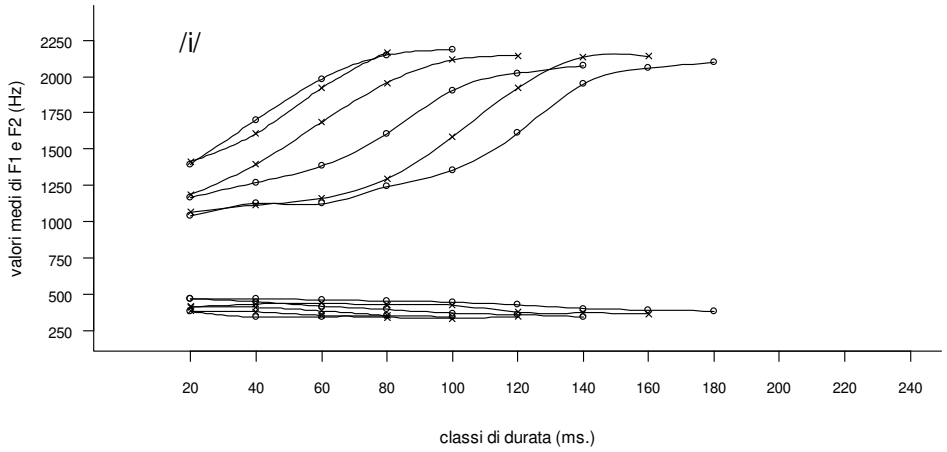


Figura 5: Traiettorie dittongali medie di /i/ distinte per classi di durata
(il grafico è stato realizzato con i dati relativi alla parola 'pesci' nel corpus di Pozzuoli)

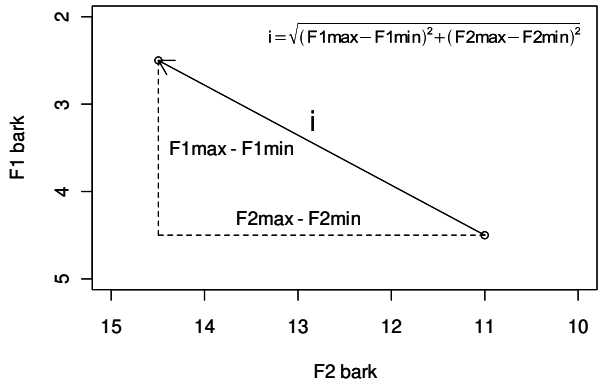


Figura 6: Il coefficiente di dittongazione

Come esemplificato in figura 6, tali misure sono rappresentabili in uno spazio F2-F1 come i cateti di un triangolo rettangolo. Pertanto, è possibile ottenere un unico indice dell'escursione dei valori delle due formanti calcolando l'ipotenusa di questo triangolo attraverso il teorema di Pitagora. La formula per il calcolo dell'indice è dunque la seguente:

$$\text{coeff. ditt.} = \sqrt{(F1max - F1min)^2 + (F2max - F2min)^2}$$

Prima di essere immessi nella formula, i valori in Hertz vengono convertiti in Bark secondo la formula di Traunmüller (1990). In tal modo viene valutato in maniera più adeguata l'apporto dei movimenti della prima formante all'ampiezza totale del movimento dittongale. I movimenti della F1, infatti, risultano meno ampi se considerati dal punto di vista della scala acustica in Hertz, mentre vengono rappresentati più adeguatamente se considerati dal punto di vista della scala uditiva in Bark.

A tale indice è stato dato il nome di 'coefficiente di dittongazione'. Il coefficiente ha valore 0 solo nei casi di segmenti vocalici più brevi di 60 ms., giacché questi vengono caratterizzati da una sola misurazione delle formanti in un punto medio. Per quanto riguarda invece vocali più lunghe di 60 ms., il coefficiente di dittongazione si mantiene in genere al disotto del valore 1 per gli esiti che percettivamente risultano di tipo monottongale, mentre supera il valore 1.8 per gli esiti che percettivamente risultano di tipo dittongale. All'ascolto impressionistico dei dati sembra che segmenti caratterizzati da un coefficiente di dittongazione inferiore a 1.8 non producano la percezione di un dittongo.[16]

5. ANALISI DEI DATI

5.1 Allungamento prepausale

L'allungamento prepausale, ossia l'allungamento di vocali e consonanti nelle posizioni finali di diversi costituenti prosodici, è un fenomeno noto in moltissime lingue e su di esso esiste una ricca bibliografia.[17] In molti degli studi riportati in letteratura tale fenomeno risulta di tipo incrementale: le durate si fanno progressivamente più lunghe a mano a mano che si sale nella gerarchia dei costituenti prosodici (cfr. ad es. Ladd & Campbell, 1991; Wightman *et al.*, 1992). Le variazioni di durata costituiscono pertanto una delle risorse fondamentali che i parlanti utilizzano per la demarcazione di strutture prosodiche. La natura plausibilmente universale dell'allungamento prepausale è da mettere in relazione con tendenze fisiologiche comuni ad ogni attività motoria di livello superiore (cfr. Vaissière, 1983 e 1995). Tuttavia, nonostante questa base fisiologica, gli effetti di questo fenomeno sono diversi da lingua a lingua, sia per quanto riguarda il dominio dell'allungamento (numero di sillabe coinvolte), sia relativamente ai segmenti interessati (vocali e consonanti), sia per il diverso influsso esercitato dai diversi costituenti prosodici.

Nel presente studio è stata analizzata la durata dei segmenti vocalici tonici in rapporto alla posizione interna o finale di alcuni costituenti prosodici (cfr. § 3). L'analisi acustica delle variazioni di durata ha evidenziato nei dialetti di Pozzuoli e Torre Annunziata gli effetti consistenti dell'allungamento prepausale. Tale allungamento si manifesta con molta

[16] Questa 'soglia critica' per la percezione di dittongazione è stata definita su base impressionistica. In futuro si prevede di approfondire meglio tale questione attraverso lo sviluppo di adeguati test percettivi.

[17] Per l'inglese Lehiste (1972); Oller (1973); Klatt (1975 e 1976); Crystal & House (1990); Edwards *et al.* (1991); Byrd (2000); per il francese Grosjean & Deschamps (1972); Crompton (1980); Rietveld (1980); Fletcher (1991); per lo svedese Lindblom (1968); Lindblom & Rapp (1973); Lyberg (1981); per l'ebraico Berkovits (1984 e 1993); per il tedesco Kohler (1983); per l'italiano o varietà regionali di italiano Vayra & Fowler (1992); Sorianello (1994 e 2006); Dell'Aglio *et al.* (2002); Bertinetto et *al.* (2006).

regolarità soprattutto nelle posizioni che hanno una più rigorosa definizione prosodica, quali la posizione finale di sintagma fonologico e la posizione finale di sintagma intonativo, mentre i dati sono più variabili e specifici di ciascuna varietà per quanto riguarda le sottocategorie della posizione finale di sintagma intonativo (SI_N, SI_Q, SI_C, SI_L), le quali sono state definite invece su base pragmatica e sulla presenza o meno di intonazione ascendente.

I segmenti vocalici tonici del dialetto di Pozzuoli sono particolarmente sensibili alla posizione nella struttura prosodica. La durata delle vocali, infatti, aumenta in corrispondenza dei confini prosodici di ordine gerarchico superiore. Come già documentato per altre lingue, tale allungamento è di tipo incrementale, cioè incide in maniera sempre maggiore quanto più si sale nella gerarchia prosodica. In tabella 2 (appendice) si riportano i dati relativi alla durata di tutti i *tokens* del corpus di Pozzuoli nelle diverse posizioni prosodiche definite per questa ricerca. Per ciascuna classe prosodica vengono riportati il valore minimo, il primo quartile, la mediana, la media, il terzo quartile[18], il valore massimo, il numero dei *tokens* analizzati.

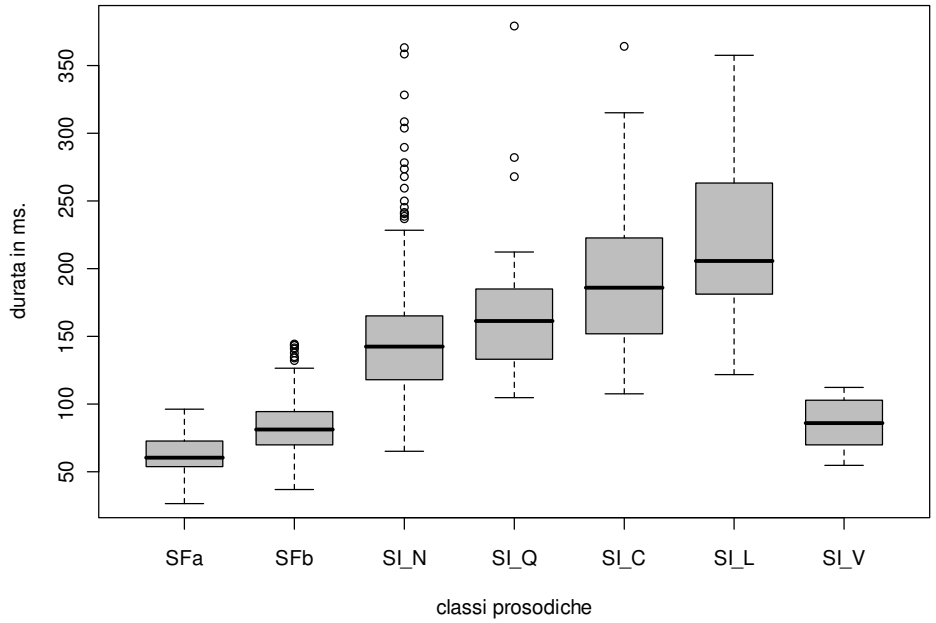


Figura 7: Durate per posizione prosodica nel corpus di Pozzuoli

I dati della tabella 2 sono riportati in un grafico del tipo *box plot* in figura 7. Il grafico evidenzia come i valori di durata in posizione finale di sintagma intonativo (SI_N) siano molto maggiori di quelli in posizione finale di sintagma fonologico (SFb) e come questi ultimi siano a loro volta sensibilmente maggiori dei valori in posizione interna di sintagma fonologico (SFa). Inoltre, a parità di posizione finale di sintagma intonativo, la durata

[18] I quartili (compresa la mediana) e le medie sono stati calcolati escludendo gli *outliers*. Gli *outliers* vengono individuati secondo la formula proposta da Tukey (1977).

continua ad aumentare progressivamente se l'enunciato ha un'intonazione di tipo interrogativo, di tipo 'continuativo', o del tipo 'lista'. Le vocali etichettate come SI_V, che pur essendo percepite in posizione finale di sintagma intonativo sembravano essere realizzate con particolare velocità, hanno in effetti durata sensibilmente inferiore alle vocali in posizione finale di sintagma intonativo e sono comparabili invece alle vocali in posizione finale di sintagma fonologico. Si è deciso di escludere da questa rappresentazione le vocali prodotte con fenomeni di esitazione,[19] in quanto i valori di durata di tali vocali risultano troppo variabili e non costituiscono in effetti una classe prosodica a sé stante.

Per valutare la significatività statistica delle differenze riscontrate tra le mediane nelle diverse classi prosodiche è stata effettuato il test non parametrico di Kruskal-Wallis.[20] Il test è stato applicato prima all'insieme dei campioni, per accertare che almeno una delle differenze tra le mediane non fosse dovuta al caso, quindi sono stati effettuati confronti appaiati tra ciascun campione con tutti gli altri. I risultati del test come valori di p. sono riportati nella tabella 3 (appendice), insieme con le differenze in ms. e in percentuale[21] tra le mediane di ogni classe prosodica con ciascuna delle altre classi. Dal test risulta che quasi tutte le differenze tra le mediane sono statisticamente significative per un valore di p. < 0.01, ad eccezione delle differenze tra SI_N e SI_Q da un lato, e SI_Q e SI_C dall'altro, per le quali p. < 0.05, ma > 0.01. Non risulta esserci una differenza statisticamente significativa tra vocali in posizione finale di sintagma fonologico (SFb) e vocali finali di sintagma intonativo realizzate velocemente (SI_V).

Le differenze di durata tra le varie posizioni prosodiche, pur essendo in genere statisticamente significative (con le eccezioni fatte sopra), non sono tutte di uguale consistenza. Confrontando i valori in tabella 3 con il grafico in figura 7, emerge come la differenza di durata più consistente sia quella tra vocali in posizione interna di sintagma intonativo (SFa e SFb) da un lato e vocali in posizione finale di sintagma intonativo dall'altro. Le vocali in posizione SI_N durano il 75 % in più di quelle in posizione SFb e il 136 % in più di quelle in posizione SFa. A parità di posizione finale di sintagma intonativo, si presenta una certa polarizzazione tra vocali non marcate (SI_N) e vocali con intonazione di 'continuazione' e del tipo 'lista' (SI_C e SI_L). Rispetto alle vocali in posizione SI_N, quelle in posizione SI_C durano il 30 % in più e quelle in posizione SI_L presentano un incremento del 44 %. Meno consistenti, invece, le differenze tra SI_N e SI_Q, SI_Q e SI_C, SI_C e SI_L.

Anche nel dialetto di Torre Annunziata la durata dei segmenti vocalici tonici è fortemente influenzata dalla posizione nella struttura prosodica. La tabella 4 (appendice) e la figura 8 riassumono i dati sulle variazioni di durata. Come per Pozzuoli, anche in questo caso l'allungamento è di tipo incrementale, cioè aumenta progressivamente nel passaggio

[19] Tali vocali sono etichettate come SI_H. Le statistiche riassuntive relative a questa categoria sono riportate nelle tabelle in appendice.

[20] Il test di Kruskal-Wallis è un'alternativa non parametrica all'ANOVA. Per una descrizione tecnica del test si veda Hollander & Wolfe (1973: 115-120). La scelta di un test non parametrico è stata dettata dalla natura dei nostri dati, che non soddisfano alcuni requisiti delle statistiche parametriche, come la normalità delle distribuzioni e l'omogeneità delle varianze.

[21] I valori in ms. e in percentuale vanno letti come incremento del valore nella colonna di destra sul valore della colonna di sinistra.

dalla posizione interna al sintagma fonologico (SFa) a quella finale di sintagma fonologico ma interna al sintagma intonativo (SFb), a quella finale di sintagma intonativo (SI_N). Inoltre, i segmenti in posizione finale di sintagma intonativo caratterizzati da intonazione ascendente (SI_Q, SI_C, SI_L) presentano durate sensibilmente maggiori di quelle dei segmenti nella stessa posizione ma privi di tale contorno intonativo (SI_N). Poco significative, invece, le differenze che si notano all'interno delle diverse classi caratterizzate da intonazione ascendente.

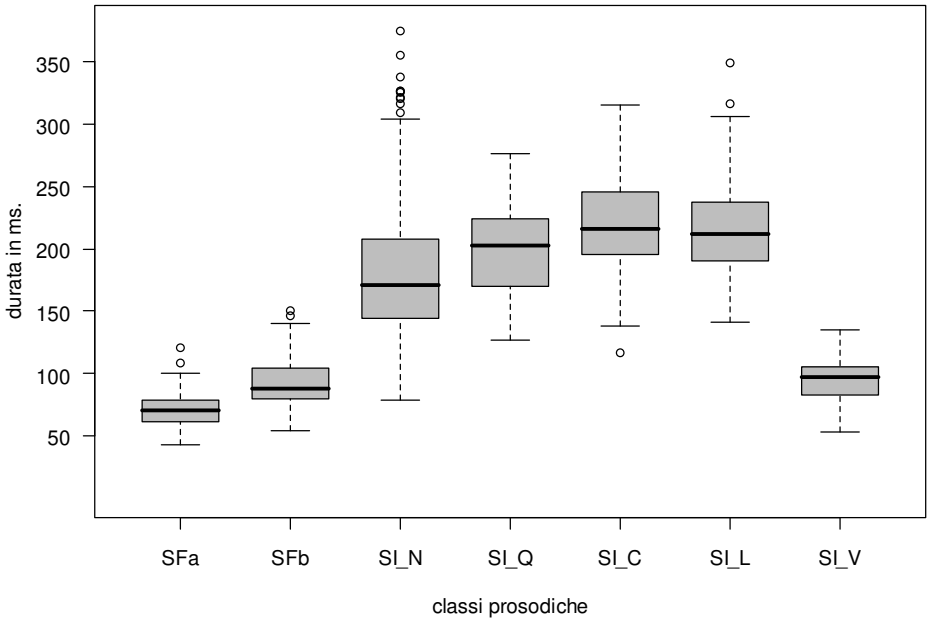


Figura 8: Durate per posizione prosodica nel corpus di Torre Annunziata

La significatività statistica delle differenze notate tra le mediane di ciascuna classe è stata valutata attraverso il test non-parametrico di Kruskal-Wallis. I risultati sono riportati come valori di p. nella tabella 5, insieme con le differenze in millisecondi e in percentuale tra le mediane. Le differenze che non risultano statisticamente significative sono quelle tra SI_N e SI_Q, SI_Q e SI_C, SI_C e SI_L. Non è stata dunque rilevata una differenza significativa tra durate in posizione finale di enunciati con intonazione di 'continuazione' da un lato, e del tipo 'lista' dall'altro, mentre nel dialetto di Pozzuoli quest'ultima classe presentava durate sensibilmente maggiori. Come per Pozzuoli, invece, le vocali in posizione finale di sintagma intonativo realizzate con particolare velocità di eloquio (SI_V) sono assimilabili a quelle in posizione finale di sintagma fonologico (SFb).

Le differenze più consistenti sono quelle che intercorrono tra segmenti in posizione non-prepausale e segmenti in posizione prepausale: le vocali in posizione finale di sintagma intonativo non marcata (SI_N) durano il 94 % in più di quelle in posizione finale di sintagma fonologico (SFb) e il 144 % in più di quelle in posizione interna al sintagma fonologico (SFa). Un incremento sensibile, di circa il 26 %, avviene anche nel passaggio dalla posizione interna al sintagma fonologico (SFa) a quella finale di sintagma fonologico

(SFb). A loro volta i segmenti in posizione finale di sintagma intonativo con intonazione di 'continuazione' (SI_C) o del tipo 'lista' (SI_L) hanno durate sensibilmente superiori a quelle dei segmenti in posizione finale di sintagma intonativo non marcata (SI_N), rispettivamente del 26 % e del 24 %.

I dati sull'allungamento prepausale delle vocali toniche ottenuti in questa ricerca presentano un quadro piuttosto simile a quello già rilevato per le toniche in posizione finale e non finale in studi sperimentali condotti su diverse varietà di italiano (cfr. Albano Leoni *et al.* 1995, Dell'Aglio *et al.* 2002, Sorianello & Calamai 2005, Sorianello 2006).[22]

5.2 Variazioni nel coefficiente di dittongazione

Come descritto in § 4, per ogni *token* vocalico è stato calcolato un coefficiente di dittongazione, che fornisce un'indicazione dell'ampiezza del movimento dittongale. Gli esiti di tipo monottongale presentano valori minimi, al disotto di 1, mentre esiti tipicamente dittongali superano in genere il valore 2 del coefficiente. L'analisi impressionistica dei dati ha indotto a considerare un coefficiente di 1.8 come soglia critica al disotto della quale le traiettorie formantiche non vengono percepite come un movimento dittongale.

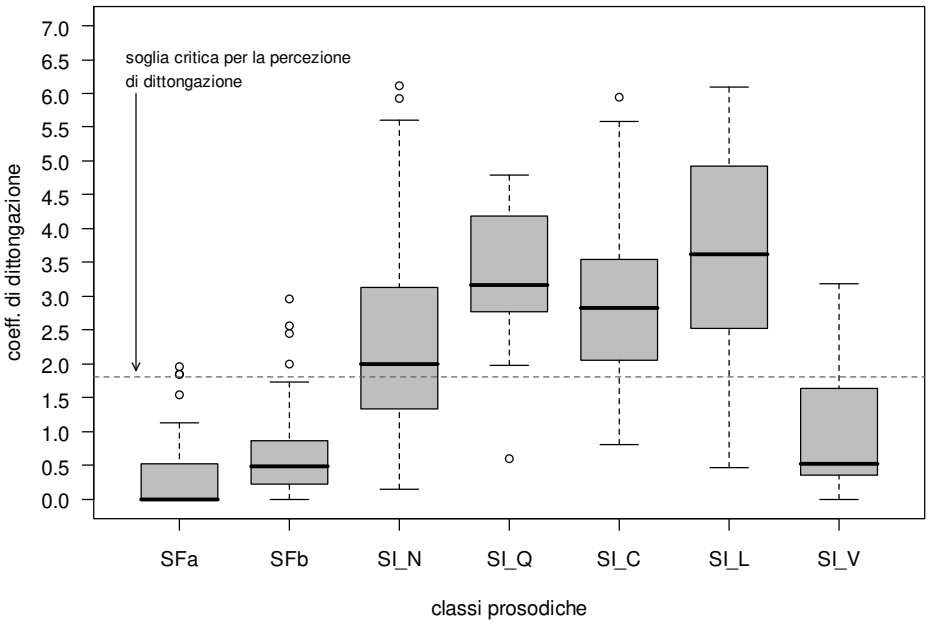


Figura 9: Coefficienti di dittongazione in rapporto alla posizione prosodica nel corpus di Pozzuoli[23]

[22] Per un confronto più agevole è necessario accorpare in un'unica posizione 'interna' i dati relativi alle posizioni SFa e SFb definite per questa ricerca.

[23] Sono stati esclusi i *tokens* di /ɛ/, che non presentano dittongazione nei parlanti più anziani (cfr. nota 8).

Il coefficiente di dittongazione permette di indagare eventuali correlazioni tra dittongazione e altri parametri di riferimento. Nella tabella 6 (appendice) si riportano dati riassuntivi sul coefficiente di dittongazione nelle diverse classi prosodiche per il dialetto di Pozzuoli. I dati sono rappresentati graficamente in figura 9.

Dalla figura appare chiaramente la stretta correlazione tra dittongazione e posizione della parola nella struttura prosodica. Gli esiti delle variabili esaminate si presentano come esclusivamente monottongali (con sporadiche eccezioni) in posizione interna e finale di sintagma fonologico, mentre sono prevalentemente dittongali in posizione finale di sintagma intonativo,[24] con movimenti formantici particolarmente ampi nelle posizioni caratterizzate da intonazione ascendente (SI_Q, SI_C, SI_L).

Come per le variazioni di durata, sono stati effettuati dei test statistici per valutare l'affidabilità delle differenze notate tra le mediane del coefficiente di dittongazione nelle varie classi prosodiche. Nella tabella 7 (appendice) sono riportati i risultati del test di Kruskal-Wallis come valori di p., insieme con le differenze in Bark e in percentuale tra le mediane di ogni classe prosodica con ciascuna delle altre classi.

Dai test emerge che le differenze nelle mediane del coefficiente di dittongazione tra SI_Q e SI_C, e SI_Q e SI_L non sono statisticamente significative, mentre la differenza per le classi SI_C e SI_L è significativa per un valore di $p. < 0.05$ ma > 0.01. Infine, non risulta significativa la differenza delle mediane nei valori del coefficiente di dittongazione tra le classi SFb e SI_V. Dalle differenze in percentuale emerge, come per le durate, una netta contrapposizione tra realizzazioni in posizione interna e finale di sintagma fonologico e le realizzazioni in posizione finale di sintagma intonativo. Gli esiti in posizione finale di sintagma intonativo non marcata (SI_N) presentano un incremento nel coefficiente di dittongazione del 308% rispetto agli esiti in posizione finale di sintagma fonologico (SFb). A parità di posizione finale di Sintagma Intonativo, solo le realizzazioni con intonazione interrogativa (SI_Q) e intonazione del tipo 'lista' (SI_L) sembrano avere coefficienti di dittongazione sensibilmente maggiori.

In tabella 8 (appendice) e in figura 10 si riportano i dati sul coefficiente di dittongazione nel dialetto di Torre Annunziata. Anche per questa varietà si evidenzia una netta polarizzazione tra esiti di tipo monottongale, caratteristici delle posizioni non-prepausali, interna e finale di sintagma fonologico (SFa e SFb), e esiti dittongali, caratteristici delle posizioni finali di sintagma intonativo (SI_N, SI_Q, SI_C, SI_L). I segmenti in posizione SI_V, cioè in posizione finale di sintagma intonativo ma realizzati con particolare velocità di eloquio, non presentano dittongazione, se non piuttosto sporadicamente. In posizione prepausale, invece, gli esiti sono prevalentemente dittongali, ma sono nondimeno presenti esiti monottongali, anche se in percentuale nettamente inferiore.

In posizione finale di sintagma intonativo non si riscontrano grosse differenze nel coefficiente di dittongazione in rapporto al tipo di andamento dell'intonazione, tranne nel caso degli enunciati caratterizzati da intonazione di 'continuazione', che presentano coefficienti di dittongazione sensibilmente maggiori. Per valutare la significatività statistica delle differenze tra le mediane dei coefficienti di dittongazione nelle varie classi prosodiche, sono stati effettuati confronti appaiati tra le mediane con il test di Kruskal-

[24] Sulla pertinenza della posizione finale di sintagma intonativo per l'emersione delle varianti dittongali si veda quanto già riportato in Loporcaro (1988: 519 ss.) per il dialetto di Altamura.

Wallis. I risultati dei test come valori di *p.* sono riportati nella tabella 10, insieme con le differenze tra le mediane in Bark e in percentuale. Dai test statistici risultano significative tutte le differenze riscontrate, ad eccezione delle differenze tra SI_N e SI_Q, SI_N e SI_L, SI_Q e SI_C, SI_Q e SI_L. In altre parole, all'interno della macrocategoria prepausale risulta statisticamente significativo soltanto l'incremento di durata che SI_C presenta rispetto a SI_N, e che SI_L presenta rispetto a SI_C; non sono significative, invece, le altre differenze.

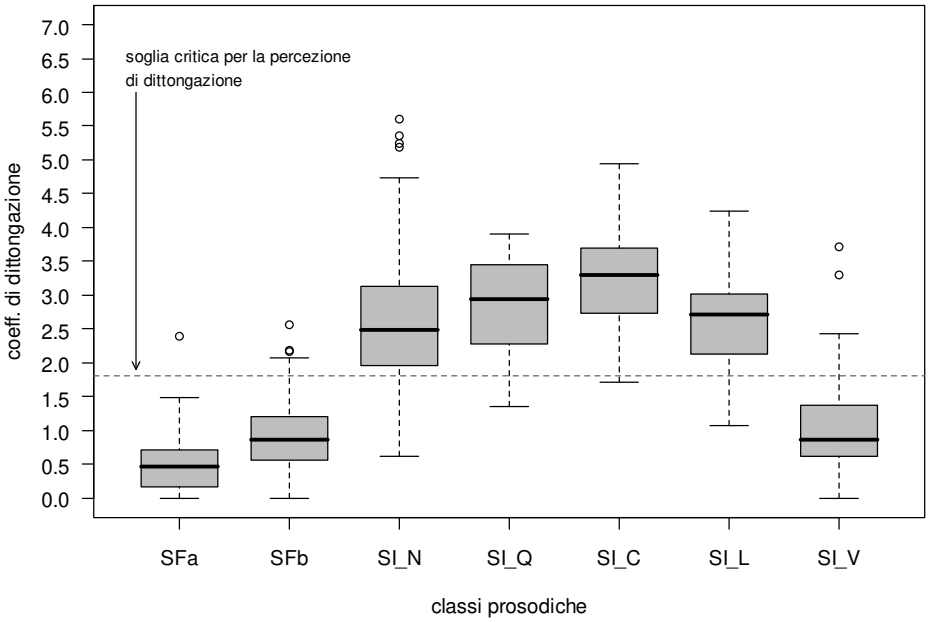


Figura 10: Coefficienti di dittongazione in rapporto alla posizione prosodica nel corpus di Torre Annunziata

Guardando gli incrementi in percentuale, si riscontra un netto aumento del coefficiente di dittongazione (186 %) nel passaggio dalla posizione finale di sintagma fonologico (SFb) a quella finale di sintagma intonativo non marcata (SI_N). Di non molto rilievo le differenze che si riscontrano tra le varie sottoclassi della posizione finale di sintagma intonativo. Il coefficiente di dittongazione in posizione finale con intonazione di 'continuazione' (SI_C) è superiore del 33 % rispetto ai valori in posizione finale non marcata (SI_N); il coefficiente di dittongazione in posizione finale con intonazione del tipo 'lista' (SI_L) è inferiore di circa il 18 % rispetto ai valori in posizione finale con intonazione di 'continuazione' (SI_C).[25] Per il resto, le differenze sono di poco conto e, come già evidenziato in precedenza, non risultano statisticamente significative.

[25] Per questo ultimo aspetto Torre Annunziata presenta una situazione inversa rispetto a quella di Pozzuoli.

6. CONCLUSIONI

In questo contributo sono stati presentati i primi risultati di un'indagine acustica condotta su un fenomeno di alternanza tra esiti monottongali e esiti dittongali di alcune variabili vocaliche in alcuni dialetti meridionali. L'analisi dei dati, limitata per il momento a Pozzuoli e Torre Annunziata, ha messo in evidenza due aspetti fondamentali di questo fenomeno, entrambi legati alla posizione delle variabili vocaliche nella struttura prosodica: l'allungamento prepausale e la tendenza delle varianti dittongali a emergere nella posizione finale di sintagma intonativo.[26]

L'analisi delle durate ha mostrato la presenza di un forte allungamento dei segmenti vocalici tonici nella posizione finale di sintagma intonativo rispetto alle durate in posizione interna (l'incremento è del 92% per Pozzuoli e del 116% per Torre Annunziata).[27] Un allungamento statisticamente significativo, anche se più modesto, è stato rilevato nella posizione finale di sintagma fonologico rispetto alle durate in posizione interna a tale costituente (l'incremento è del 35% per Pozzuoli e del 26% per Torre Annunziata). I dati relativi alle durate nelle sottocategorie definite per la posizione finale di sintagma intonativo sono meno coerenti, ma si nota comunque un allungamento significativo negli enunciati caratterizzati da intonazione ascendente. Come già documentato in diversi studi (cfr. ad es. Ladd & Campbell, 1991; Wightman *et alii*, 1992), l'allungamento è dunque di tipo incrementale: le durate si fanno progressivamente più lunghe a mano a mano che si sale nella gerarchia dei costituenti prosodici.

L'analisi del coefficiente di dittongazione ha evidenziato una polarizzazione piuttosto netta tra le realizzazioni della variabili vocaliche in posizione interna e quelle in posizione finale di sintagma intonativo. Mentre in posizione interna il coefficiente si mantiene su livelli piuttosto bassi, in posizione finale esso supera tendenzialmente (e spesso in maniera consistente) il valore di 1.8, che su base impressionistica è stato individuato come soglia critica per la percezione di dittongazione. L'alternanza monottongo/dittongo in questi dialetti rientra quindi tra i fenomeni di variazione fonetica che dipendono dalla posizione nella struttura prosodica e che forniscono al parlante indici acustici per la segmentazione della catena palata in costituenti prosodici.

Rispetto al quadro delineato sono presenti comunque anche dati in controtendenza. In particolare si notano diversi esiti monottongali in posizione finale, laddove sarebbero stati attesi esiti dittongali. A questi proposito occorre precisare che nei dialetti esaminati l'alternanza monottongo/dittongo non è regolata soltanto da fattori strutturali interni, ma assume anche significati sociolinguistici. Assodato che i dittonghi sono in genere esclusi da posizioni interne, nei contesti favorevoli alla dittongazione le varianti dittongali sono spesso in competizione con varianti monottongali. Tale competizione è intrisa di valori extralinguistici, in quanto inevitabilmente le varianti dittongali risultano marcate sia dal punto di vista geografico, perché assenti nella maggioranza delle comunità limitrofe, sia dal

[26] Posizione prosodica, durata e dittongazione sono strettamente interrelati. Sugli eventuali rapporti gerarchici tra queste variabili si indagherà in futuri lavori, attraverso modelli di regressione multipla che non è stato possibile esporre in questo contributo (cfr. Abete, in preparazione).

[27] Queste percentuali si ottengono includendo nella categoria 'interna' le durate relative alle posizioni SFa (interna al sintagma fonologico) e SFb (finale di sintagma fonologico ma interna al sintagma intonativo).

punto di vista diastratico, perché assenti nella lingua standard. I dittonghi, quindi, non assolvono una funzione esclusivamente 'grammaticale', marcando ad esempio la fine del sintagma intonativo, ma veicolano anche significati sociali, come l'appartenenza a un gruppo e la definizione della propria identità e della propria posizione nella società.[28] Privata della dimensione sociale, una certa variabilità tra realizzazioni dittongali e realizzazioni monottongali in fine enunciato risulta assolutamente caotica e priva di significato. Una parte dei dati in controtendenza potrebbe spiegarsi, dunque, chiamando in causa fattori sociolinguistici del tipo a cui si è accennato, ma che al momento non sono stati indagati sistematicamente. L'analisi qualitativa di queste 'eccezioni' potrebbe essere in futuro un campo di indagine molto fruttuoso, che permetterebbe di integrare nello studio dell'alternanza monottongo/dittongo anche fattori di tipo esterno.

RINGRAZIAMENTI

Si ringraziano i tre revisori anonimi, i cui commenti sono stati molto utili per la stesura finale di questo articolo.

[28] Per questi aspetti si veda Abete & Simpson (in stampa).

7. BIBLIOGRAFIA

Abete, G. (2006), Sulla questione della sillaba superpesante: i dittonghi discendenti in sillaba chiusa nel dialetto di Pozzuoli, in *Analisi prosodica. Teorie, modelli e sistemi di annotazione*, Atti del 2° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Salerno, 30 novembre – 2 dicembre 2005 (R. Savy & C. Crocco, editors), Torriana: EDK Editore, 379-398.

Abete, G. (in preparazione), *I processi di dittongazione nei dialetti dell'Italia meridionale. Un approccio sperimentale*, Tesi di dottorato, Jena: Friedrich-Schiller-Universität.

Abete, G. & Simpson, A. (in stampa), L'estensione della dittongazione nei giovani pescatori di Pozzuoli (NA). Dati acustici su un cambiamento fonetico in corso, in *La comunicazione parlata*, Atti del 3° Convegno Internazionale, Napoli, 23-25 febbraio 2009.

Albano Leoni, F., Cutugno, F. & Savy, R. (1995), The vowel system of Italian connected speech, in *Proceedings of the 13th International Congress of Phonetic Sciences* (K. Elenius & P. Branderud, editors), Stockholm, Stockholm University, vol. 4, 396-399.

Beckman, M. E. (1996), The parsing of prosody, *Language and Cognitive Processes*, 11, 17-67.

Beckman, M. E. & Pierrehumbert, J. (1986), Intonational structure in Japanese and English, *Phonology Yearbook*, 3, 255-309.

Berkovits, R. (1984), Duration and fundamental frequency in sentence final intonation, *Journal of Phonetics*, 12, 255-265.

Berkovits, R. (1993), Utterance-final lengthening and the duration of final-stop closures, *Journal of Phonetics*, 21, 479-489.

Bertinetto, P.M., Dell'Aglio, M. & Agonigi, M. (2006), Quali fattori influenzano maggiormente la durata vocalica e consonantica in italiano? Un'indagine mediante l'algoritmo di decisione C&RT, in *La Comunicazione Parlata*, Atti del congresso internazionale, Napoli 23-25 febbraio 2006 (M. Pettorino, A. Giannini, M. Vallone & R. Savy, editors), Napoli: Liguori, 13-38.

Byrd, D. (2000), Articulatory vowel lengthening and coordination at phrasal junctures, *Phonetica*, 57, 3-16.

Calamai, S., Marotta, G. & Sardelli, E. (2003), La modulazione di frequenza in due varietà toscane (Pisa e Firenze). Una indagine preliminare, *Quaderni del Laboratorio di Linguistica della Scuola Normale Superiore di Pisa*, 4 n.s., 11-25.

Cho, T. (2004), Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English, *Journal of Phonetics*, 32, 141-176.

Cho, T., McQueen, J.M. & Cox, E.A. (2007), Prosodically driven phonetic detail in speech processing: the case of domain-initial strengthening in English, *Journal of Phonetics*, 35, 210-243.

Como, P. (2006), Elicitation techniques for spoken discourse, in *Encyclopedia of language and linguistics* (K. Brown, editor), second edition, vol. 4., Amsterdam: Elsevier, 105-109.

Crompton, A. (1980), Timing patterns in French, *Phonetica*, 37, 205-234.

Crystal, T. & House, A. S. (1990), Articulation rate and the duration of syllables and stress groups in connected speech, *Journal of Acoustical Society of America*, 88 (1), 101-112.

Dell'Aglio, M., Agonigi, M. M. & Bertinetto, P. M. (2002), Le durate dei foni vocalici in rapporto al contesto nel parlato di locutori pisani. Primi risultati, in *La fonetica acustica come strumento di analisi della variazione linguistica in Italia*, Atti delle VIIe Giornate di Studio del Gruppo di Fonetica Sperimentale, Napoli, 14-15 novembre 1996 (a cura di A. Regnicoli), Roma: Il Calamo, 53-58.

Edwards, J.E., Beckman, M.E. & Fletcher, J. (1991), The articulatory kinematics of final lengthening, *Journal of the Acoustical Society of America*, 89, 369-382.

Fletcher, J. (1991), Rhythm and final lengthening in French, *Journal of Phonetics*, 19, 193-212.

Fougeron, C. & Keating, P. A. (1997), Articulatory strengthening at edges of prosodic domains, *Journal of the Acoustical Society of America*, 101 (6), 3728-3740.

Grosjean, F. & Deschamps, A. (1972), Analyse des variables temporelles du français spontanée, *Phonetica*, 26, 129-156.

Holbrook, A. & Fairbanks, G. (1962), Diphthong formants and their movements, *Journal of Speech and Hearing Research*, 5, 38-58.

Hollander, M. & Wolfe, D. A. (1973), *Nonparametric Statistical Methods*, New York: John Wiley & Sons.

Keating, P., Cho, T., Fougeron, C. & Hsu, C. (2003), Domain-initial articulatory strengthening in four languages, in *Phonetic interpretation* (J. Local, R. Ogden, R. Temple, editors), Papers in Laboratory Phonology 6, Cambridge: Cambridge University Press, 143-161.

Klatt, D. H. (1975), Vowel lengthening is syntactically determined in a connected discourse, *Journal of Phonetics*, 3, 129-140.

Klatt, D. H. (1976), Linguistic uses of segmental duration in English: acoustic and perceptual evidence, *Journal of the Acoustical Society of America*, 59, 1208-1221.

Kohler, K. J. (1983), Prosodic boundary signals in German, *Phonetica*, 40, 89-134.

Labov, W. (1972), Some principles of linguistic methodology, *Language in Society*, 1, 97-120.

Ladd, R. & Campbell, N. (1991), Theories of prosodic structure: Evidence from syllable duration, in *Proceedings of the 12th International Congress of Phonetic Sciences,* Aix-en-Provence, II, 290-293.

Lausberg, H. (1939), *Die Mundarten Südlukaniens* (Beiheft XC zur «Zeitschrift für romanische Philologie»), Halle: Niemeyer.

Lehiste, I. (1972), The timing of utterance and linguistic boundaries, *Journal of the Acoustical Society of America*, 51, 2018-2024.

Lindblom, B. (1968), Temporal organization of syllable production, *Speech Transm. Lab., Q. Prog. Status Rep*., No. 2-3, 1-5.

Lindblom, B. & Rapp, K. (1973), Some temporal regularities of spoken Swedish, *Papers in Linguistics, University of Stokholm*, 21, 1-59.

Loporcaro, M. (1988), *Grammatica storica del dialetto di Altamura*, Pisa: Giardini.

Lyberg, B. (1981), Some observations on the vowel duration and the fundamental frequency contour in Swedish utterances, *Journal of Phonetics*, 9, 261-272.

Marotta, G., Calamai, S. & Sardelli, E. (2004), Non di sola lunghezza. La modulazione di f0 come indice sociofonetico, in *Costituzione, gestione e restauro di corpora vocali*, Atti delle XIVe Giornate di Studio del G.F.S., Università della Tuscia (Viterbo), 4-6 dicembre 2003 (A. De Dominicis, L. Mori & M. Stefani, editors), Roma: Esagrafica, 215-220.

Milroy, L. (1987), *Observing and analysing natural language*, Cambridge: Blackwell.

Nearey, T. M. & Assman, P. F. (1986), Modeling the role of inherent spectral change in vowel identification, *Journal of the Acoustical Society of America*, 80, 1297-1308.

Nespor, M. (1993), *Fonologia*, Bologna: Il Mulino.

Nespor, M. & Vogel, I. (1986), *Prosodic phonology*, Dordrecht: Foris.

Oller, D. K. (1973), The effect of position in utterance on speech segment duration in English, *Journal of the Acoustical Society of America*, 54 (5), 1235-1247.

Pierrehumbert, J. & Beckman, M. E. (1988), *Japanese tone structure*, Cambridge (MA): MIT Press.

R Core Development Team (2008), *R: a language and environment for statistical computing*, http://www.R-project.org.

Rietveld, A. C. M. (1980), Word boundaries in French language, *Language and Speech*, 23(3), 289-296.

Rohlfs, G. (1938), Der Einfluß des Satzakzentes auf den Lautwandel, *Archiv für das Studium der neueren Sprachen*, CLXXIV: 54-6.

Rohlfs, G. (1966) [1949], *Grammatica storica della lingua italiana e dei suoi dialetti*, 1: Fonetica, Torino: Einaudi.

Selkirk, E. (1984), *Phonology and syntax: The relation between sound and structure*, Cambridge (MA): MIT Press.

Selkirk, E. (1986), On derived domains in sentence phonology, *Phonology Yearbook*, 3, 371-405.

Shattuck-Hufnagel, S. & Turk, A. E. (1996), A prosody tutorial for investigators of auditory sentence processing, *Journal of Psycholinguistic Research*, 25, 193-247.

Simpson, A. P. (1998), Characterizing the formant movements of German diphthongs in spontaneous speech, in *Computer Linguistik und Phonetik zwischen Sprache und Sprechen*, Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache – KONVENS – 98, Bonn, 5-7 ottobre 1998 (B. Schröder, W. Lenders, W. Hess & T. Portele, editors), Frankfurt: Lang, 192-200.

Sjölander, K. & Beskow, J. (2006), *Wavesurfer*, Department of Speech and Music Communication, Stockholm: KTH.

Sorianello, P. (1994), Il processo dell'allungamento prepausale: dati ed interpretazioni, *Quaderni del dipartimento di Linguistica, Università di Firenze*, 5, 47-73.

Sorianello, P. (2006), Per una definizione fonetica e fonologica dei confini prosodici, in *La Comunicazione Parlata*, Atti del convegno internazionale di studi, Napoli 23-25 febbraio 2006 (M. Pettorino, A. Giannini, M. Vallone, R. Savy, editors), Napoli: Liguori, 298-318.

Sorianello, P. & Calamai, S. (2005), Il sistema vocalico romano, in *Italiano parlato. Analisi di un dialogo* (F. Albano Leoni & R. Giordano, editors), Napoli: Liguori, 25-70.

Traunmüller, H. (1990), Analytical expressions for the tonotopic sensory scale, *Journal of the Acoustical Society of America*, 88, 97-100.

Tukey, J. W. (1977), *Exploratory data analysis*, Reading: Addison-Wesley .

Vaissière, J. (1983), Language-independent prosodic features, in *Prosody: models and measurements* (A. Cutler & D. R. Ladd, editors), Berlin: Springer, 53-66.

Vaissière, J. (1995), Phonetic explanations for cross-linguistic prosodic similarities, *Phonetica*, 52, 123-130.

Vayra, M. & Fowler, C. (1992), Declination of supralaryngeal gestures in spoken Italian, *Phonetica*, 49(1), 48-60.

Wightman, C. W., Shattuk-Hufnagel, S., Ostendorf, M. & Price, P. J. (1992), Segmental durations in the vicinity of prosodic phrase boundaries, *Journal of the Acoustical Society of America*, 91, 1707-1717.

1. APPENDICE: DATI RIASSUNTIVI DELLE ANALISI EFFETTUATE E RISULTATI DEI TEST STATISTICI

	SFa	SFb	SI_N	SI_Q	SI_H	SI_C	SI_L	SI_V
Min.	26,00	36,00	65,00	104,00	85,00	107,00	121,00	54,00
I Qu.	53,00	69,00	118,00	133,50	114,20	153,00	181,00	69,50
Mediana	**60,00**	**81,00**	**142,00**	**161,00**	**160,00**	**186,00**	**205,00**	**85,00**
Media	60,76	82,30	148,70	172,20	184,30	190,20	220,70	86,52
III Qu.	71,75	94,00	165,00	183,00	260,20	221,50	263,00	102,50
Max.	96,00	144,00	424,00	379,00	330,00	364,00	446,00	112,00
tokens	122	259	352	24	30	80	53	27

Tabella 2 Dati riassuntivi della durata in rapporto alla posizione prosodica per il corpus di Pozzuoli

	SFa						
ms.	21	**SFb**					
%	35						
p	< 0.01						
ms.	82	61	**SI_N**				
%	136,7	75,3					
p	< 0.01	< 0.01					
ms.	101	80	19	**SI_Q**			
%	168,3	98,8	13,4				
p	< 0.01	< 0.01	**0.039**				
ms.	126	105	44	25	**SI_C**		
%	210	129,6	31	15,5			
p	< 0.01	< 0.01	< 0.01	**0.028**			
ms.	145	124	63	44	19	**SI_L**	
%	241,7	153,1	44,4	27,3	10,2		
p	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01		
ms.	25	4	-57	-76	-101	-120	**SI_V**
%	41,7	4,9	-40,1	-47,2	-54,3	-58,5	
p	< 0.01	**> 0.05**	< 0.01	< 0.01	< 0.01	< 0.01	

Tabella 3: Differenze delle mediane delle durate nelle diverse classi prosodiche (Pozzuoli). Valori in ms., in percentuale, e valori di p. nel test di Kruskal-Wallis (*One-Way ANOVA by Ranks*). Il risultato del test applicato all'insieme dei campioni (preliminarmente ai confronti appaiati) è: $\chi^2 = 657.921$, df = 6, p. < 2.2e-16

	SFa	SFb	SI_N	SI_Q	SI_H	SI_C	SI_L	SI_V
Min.	43,00	54,00	79,00	176,00	127,00	117,00	141,00	53,00
I Qu.	61,50	80,00	144,00	213,50	170,50	195,50	190,50	83,25
Mediana	**70,00**	**88,00**	**171,00**	**244,00**	**203,00**	**216,00**	**211,50**	**97,50**
Media	70,86	92,43	180,10	252,40	203,80	221,40	222,70	94,31
III Qu.	79,00	4,00	208,00	279,00	222,80	245,50	237,50	5,00
Max.	21,00	50,00	375,00	362,00	277,00	315,00	402,00	35,00
tokens	135	133	392	12	7	43	42	32

Tabella 4: Dati riassuntivi della durata in rapporto alla posizione prosodica per il corpus di Torre Annunziata

	SFa						
ms.	18	**SFb**					
%	25,7						
p	< 0.01						
ms.	101	83	**SI_N**				
%	144,3	94,3					
p	< 0.01	< 0.01					
ms.	174	156	73	**SI_Q**			
%	248,6	177,3	42,7				
p	< 0.01	< 0.01	**0.0501**				
ms.	146	128	45	-28	**SI_C**		
%	208,6	145,4	26,3	-11,5			
p	< 0.01	< 0.01	< 0.01	**> 0.05**			
ms.	141,5	123,5	40,5	-32,5	-4,5	**SI_L**	
%	202,1	140,3	23,7	-13,3	-2,1		
p	< 0.01	< 0.01	< 0.01	**> 0.05**	**> 0.05**		
ms.	27,5	9,5	-73,5	-146,5	-118,5	-114	**SI_V**
%	39,3	10,8	-43	-60	-54,9	-54	
p	< 0.01	**> 0.05**	< 0.01	< 0.01	< 0.01	< 0.01	

Tabella 5: Differenze delle mediane delle durate nelle diverse classi prosodiche (Torre Annunziata). Valori in ms., in percentuale, e valori di p. nel test di Kruskal-Wallis (*One-Way ANOVA by Ranks*). Il risultato del test applicato all'insieme dei campioni (preliminarmente ai confronti appaiati) è: $\chi^2 = 572.1547$, df = 6 , p. < 2.2e-16.

	SFa	SFb	SI_N	SI_Q	SI_H	SI_C	SI_L	SI_V
Min.	0,00	0,00	0,15	0,60	0,65	0,80	0,47	0,00
I Qu.	0,00	0,22	1,34	2,77	1,05	2,09	2,57	0,35
Mediana	**0,00**	**0,49**	**2,00**	**3,16**	**2,48**	**2,84**	**3,62**	**0,52**
Media	0,30	0,59	2,34	3,28	2,34	2,93	3,55	1,04
III Qu.	0,52	0,86	3,13	4,18	3,44	3,52	4,87	1,64
Max.	1,95	2,96	6,11	4,80	4,23	5,95	6,09	3,18
tokens	107	154	183	17	15	62	32	21

Tabella 6: Dati riassuntivi sul coefficiente di dittongazione in rapporto alla posizione prosodica per il corpus di Pozzuoli.

	SFa						
bark	0,49	**SFb**					
%	//						
p	< 0.01						
bark	2	1,51	**SI_N**				
%	//	308					
p	< 0.01	< 0.01					
bark	3,16	2,67	1,16	**SI_Q**			
%	//	545	58				
p	< 0.01	< 0.01	< 0.01				
bark	2,84	2,35	0,84	-0,32	**SI_C**		
%	//	480	42	-10			
p	< 0.01	< 0.01	< 0.01	**> 0.05**			
bark	3,62	3,13	1,62	0,46	0,78	**SI_L**	
%	//	639	81	15	27		
p	< 0.01	< 0.01	< 0.01	**> 0.05**	**< 0.05**		
bark	0,52	0,03	-1,48	-2,64	-2,32	-3,1	**SI_V**
%	//	6	-74	-84	-82	-86	
p	< 0.01	**> 0.05**	< 0.01	< 0.01	< 0.01	< 0.01	

Tabella 7: Differenze delle mediane del coefficiente di dittongazione nelle diverse classi prosodiche (Pozzuoli). Valori in Bark, in percentuale, e valori di p. nel test di Kruskal-Wallis (*One-Way ANOVA by Ranks*). Il risultato del test applicato all'insieme dei campioni (preliminarmente ai confronti appaiati) è: $\chi^2 = 366.1515$, df = 6, p. < 2.2e-16

	SFa	SFb	SI_N	SI_Q	SI_H	SI_C	SI_L	SI_V
Min.	0,00	0,00	0,61	1,35	1,65	1,71	1,08	0,00
I Qu.	0,17	0,56	1,96	2,31	1,95	2,73	2,12	0,63
Mediana	**0,46**	**0,87**	**2,49**	**2,94**	**1,99**	**3,31**	**2,72**	**0,86**
Media	0,48	0,92	2,58	2,86	2,02	3,24	2,63	1,08
III Qu.	0,71	1,20	3,13	3,41	2,01	3,66	3,01	1,34
Max.	2,39	2,57	5,61	3,90	2,48	4,94	4,25	3,72
tokens	131	121	385	5	12	40	39	32

Tabella 8: Dati riassuntivi sul coefficiente di dittongazione in rapporto alla posizione prosodica par il corpus di Torre Annunziata

	SFa						
bark	0,41	**SFb**					
%	89						
p	< 0.01						
bark	2,03	1,62	**SI_N**				
%	441	186					
p	< 0.01	< 0.01					
bark	2,48	2,07	0,45	**SI_Q**			
%	539	238	18				
p	< 0.01	< 0.01	**> 0.05**				
bark	2,85	2,44	0,82	0,37	**SI_C**		
%	618	280	33	12			
p	< 0.01	< 0.01	< 0.01	**> 0.05**			
bark	2,26	1,85	0,23	-0,22	-0,59	**SI_L**	
%	491	213	9	-7	-18		
p	< 0.01	< 0.01	**> 0.05**	**> 0.05**	< 0.01		
bark	0,40	-0,02	-1,64	-2,09	-2,45	-1,87	**SI_V**
%	86	-2	-66	-71	-74	-69	
p	< 0.01	**> 0.05**	< 0.01	< 0.01	< 0.01	< 0.01	

Tabella 9: Differenze delle mediane del coefficiente di dittongazione nelle diverse classi prosodiche (Torre Annunziata). Valori in Bark, in percentuale, e valori di p. nel test di Kruskal-Wallis (*One-Way ANOVA by Ranks*). Il risultato del test applicato all'insieme dei campioni (preliminarmente ai confronti appaiati) è: $\chi^2 = 488.3946$, df = 6, p. < 2.2e-16.

PHONETIC DETAIL IN INTONATION CONTOUR DYNAMICS

Francesco Cangemi
Laboratoire Parole et Langage – Université de Provence
francesco.cangemi@lpl-aix.fr

1. ABSTRACT

The Autosegmental-Metrical theory of intonation investigates the relationship between f0 contours and post-lexical meaning. Phonetic data are represented in the phonology as a sequence of discrete, local events. The properties of the *transitions* between one event and the next are considered to be phonologically irrelevant (§ 2).

We present data on Neapolitan Italian which show a significant correlation between the shape of these transitions and the pragmatic context in which a sentence is uttered. This correlation is stronger than the one displayed by traditional autosegmental-metrical indices (§ 3 and § 4).

In the conclusions, we discuss the usefulness of our findings as a step towards the fine-tuning of the autosegmental-metrical theory (§ 5).

2. INTRODUCTION

Intonational phonology aims at describing how phonetic suprasegmental features convey post-lexical meaning in a linguistically structured way (see Ladd, 2008 for an introduction). For instance, vocal fold vibration rate data (*f0*) are used to describe acoustic differences of the same utterance in different pragmatic contexts, as in the opposition between assertive and questioning modality.

In the frame of the autosegmental-metrical theory of intonation (*AM*), phonetic (continuous) f0 data are translated into a phonological (discrete) inventory of tunes, composed by combining only two tones, high (*H*) and low (*L*). Intonation contours consist of a string of tonal events, linked to the prosodic structure of the sentence. Some tonal events, mainly the *pitch accents* (i.e. those associated to prominent syllables), can phonetically appear as a rise (or a fall) in the f0 curve. In these cases, they are analyzed as the succession of two tones (L H for rises, H L for falls).[1]

In AM, the f0 path between the two tones which compose a rising pitch accent is not regarded as phonologically relevant. Speech synthesis systems based on this framework (e.g. Pierrehumbert, 1981, Anderson *et al.*, 1984, Black & Hunt, 1996) use a simple monotonic interpolation between the two tones. Nonetheless, data from Neapolitan Italian (*NI*, D'Imperio *et al.* 2008) show that, in different pragmatic contexts, the intonation contour of the same segmental string also differs systematically in terms of the f0 path between the two tones. The curve seems to follow a concave or convex[2] path, depending on the pragmatic context in which the sentences are uttered (see Figure 1).

[1] See D'Imperio (1999) for Neapolitan Italian. Similar treatments have been proposed also for Spanish (Hualde, 2000; Face, 2001) and English (Ladd & Schepman, 2003).
[2] Note that the attributes of concave and convex refer to the half-plane above the curve.

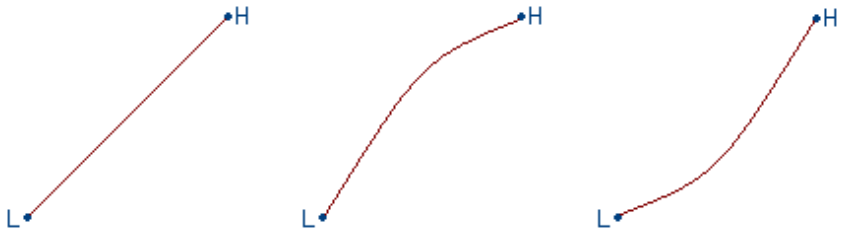


Figure 1: Sketch of linear, concave and convex interpolation

Figures 2 and 3 display the spectrogram and the f0 contour for the sentence *Milena lo vuole amaro* "Milena drinks her coffee unsweetened". In the first case, the sentence is uttered as a statement, while in the second it is uttered as a question. The most striking difference between the two f0 contours is visible in the movement associated to the last stressed syllable of the sentence (*aMAro*, highlighted by the box in both figures). In Fig. 2 we find a gradual fall, while in Fig. 3 we find a slight rise followed by a quite rapid fall. In other words, the f0 peak (H in the figures) occurs slightly before the vowel onset in the sentence, but is found later (vowel-internal) in the question, where is also visibly higher. Following the usual terminology, the H belonging to the last pitch accent is 'aligned' (in time) and 'scaled' (in frequency) differently in the two contexts.

Tone alignment and scaling are the indices usually employed in AM to define the phonetic properties of different phonological entities (e.g., of different pitch accents). But if we concentrate on the intonation contour of the first word in the sentence (*Milena*, isolated from the rest of the utterance by the vertical line in Figures 2 and 3), we notice that the rising movement associated with the stressed syllable has a different shape in the two contexts. This difference, though, does not seem to be related either to the alignment or to the scaling of the two tones: both Ls are in the first half of the stressed syllable onset, and around 225 Hz; both Hs are at the end of the stressed syllable nucleus, and around 350 Hz.

This work aims to investigate some acoustic differences considered in the AM framework as phonetic detail without phonological relevance, such as dynamic (i.e. in shape) differences. Their efficacy as indices to differentiate pragmatic contexts will be compared to that of traditional cues such as the alignment and scaling of the tones which compose the first nuclear accent of our sentences.[3] In the conclusions we will discuss the implications of our results with respect to a fine-tuning of the existing phonological model.

[3] For a discussion about the nuclearity of the first pitch accent in the (partial topic) statement utterances, see D'Imperio & Cangemi (2009).

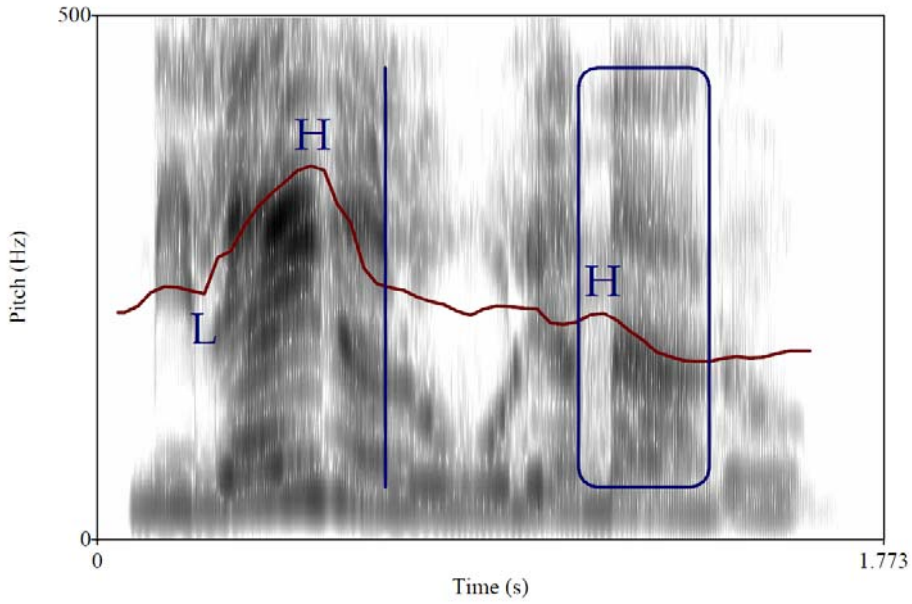


Figure 2: Utterance in (partial topic) statement context

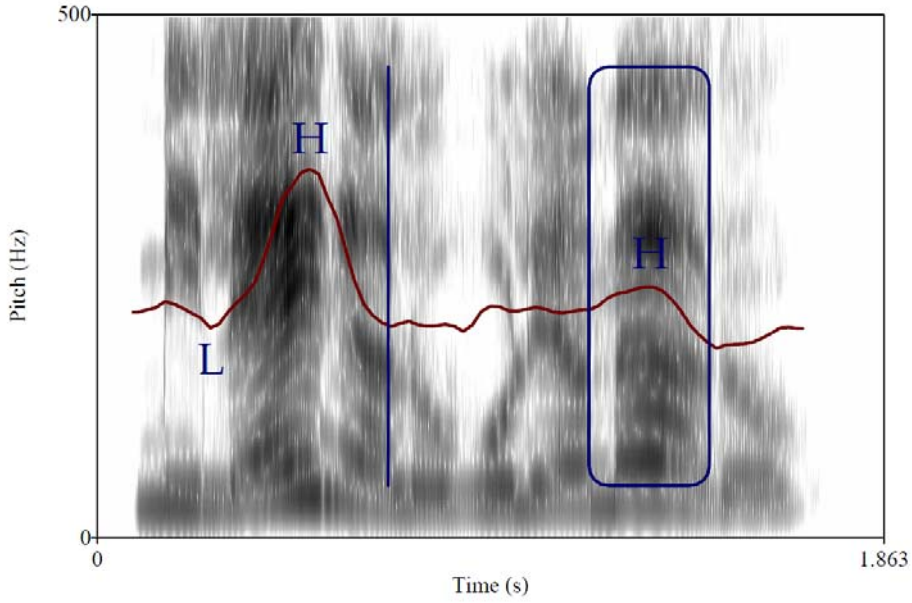


Figure 3: Utterance in (narrow focus) question context

3. MATERIALS AND METHOD

3.1 Corpus

For our study we used a subset of the corpus described in (D'Imperio *et al.*, 2008). Three native speakers of Neapolitan Italian read 30 experimental stimuli and 70 fillers in a silent room. The stimuli consisted of five repetitions of three sentences designed without voiceless plosives, which were semantically plausible and syntactically quite similar: *Amelia dorme da nonna* "Amelia sleeps at grandma's", *Valeria viene alle nove* "Valeria arrives at 9" and *Milena lo vuole amaro* "Milena wants her coffee unsweetened". The target words were all feminine proper names, agents, subjects, trisillabic, and paroxitones, with the same syllabic structure (CV) for the tonic syllable and the same quality for its nucleus (/ɛ/). The sentences were presented together with a context paragraph, which had to be read silently; this made possible the elicitation of every sentence with two different pragmatic meanings. For example, the sentence *Milena lo vuole amaro* would be interpreted (and uttered) by speakers as a Narrow Focus Question (*QNF*, meaning "Is it Milena, the one who drinks unsweetened coffee?", see Figure 3, {audio 1}) if preceded by the context:

> *After a family lunch, you're preparing coffee. You know that one of your cousins is on a diet and stays away from sugar, but you don't remember which one. You ask your aunt:...*

On the other hand, sentences preceded by the context:

> *In the afternoon, among friends, your brother is preparing coffee. He asks you whether your friends would like it sweetened or not. You don't know everybody's preferences, but only your girlfriend's. You answer:...*

would be interpreted (and uttered) as 'Partial Topic Statements' (*SPT*, meaning "As for Milena, she drinks it unsweetened; as for the others, I couldn't tell", see Figure 2, {audio 2}).[4]

The experimental material consisted of 3 subjects x 3 sentences x 2 pragmatic contexts x 5 repetitions = 90 items in total.

3.2 Measures

Target words were manually labelled in syllables using PRAAT (Boersma & Weenink, 2009). The stressed syllable, which always had a CV structure, was also labelled in segments: the labels were *Os* for the beginning of the onset (and of the entire syllable), *Ns* for the beginning of the nucleus (or the end of the onset) and *Ne* for the end of the nucleus (and of the entire syllable).[5]

The rising f0 movement in the stressed syllable was characterized by measuring the height (in Hz) and the position in time of its starting and ending points (L and H).[6] Hs were located at f0 maxima inside the stressed vowels, while the detection of Ls proved more challenging. A widely used automatic procedure is based on the detection of the local minima in the stressed syllable's onset, but we found this method too sensitive to microprosodic perturbations, which were irrelevant for our analysis. We determined that another strategy for the detection of Ls, the *two lines fitting* used for example in (D'Imperio, 2000), was not suited for our goals.

[4] For the notion of 'partial topic', see Büring (1997).
[5] See Figure 4: Os, Ns and Ne on x-axis.
[6] See Figure 4: y(L) and y(H) on y-axis for height, and x(L) and x(H) on x-axis for position.

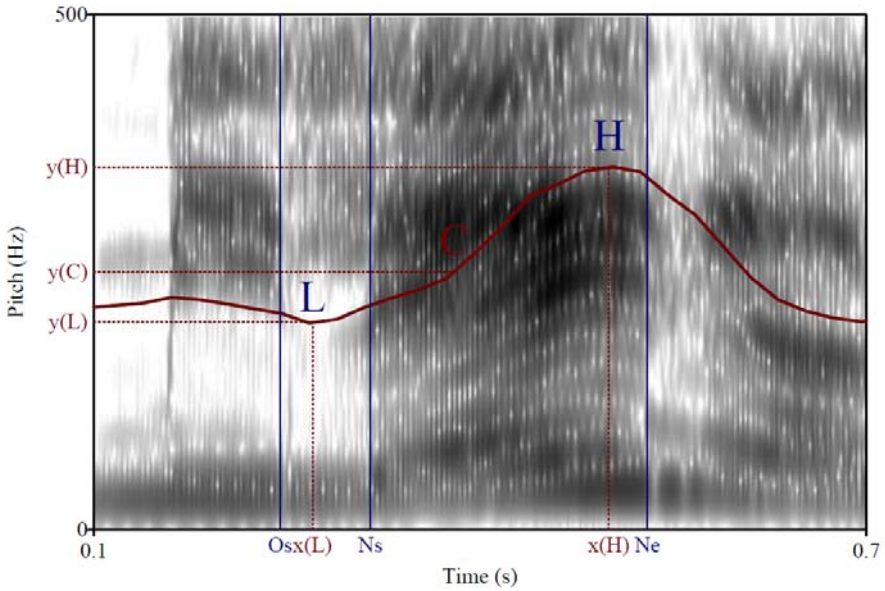


Figure 4: Measures

With this technique, the region in which the L must be found (in our case, the f0 stretch from utterance start to H) is divided into steps. For each point, two straight lines are fitted with a linear regression to the contour on its left and on its right. The L is chosen as the point associated with the pair of lines leading to the smallest modelling error. Since the differences between a concave and a convex rise have consequences on modelling and errors, the algorithm often locates Ls away from the *elbow*, the point in which the f0 curve visibly bends upwards. Concave shapes tend to be associated to an L on the left of the real elbow, and for convex ones the L is detected on its right. This means of course that we would still have an index to express our differences in interpolation, but in this case the information is conveyed in an implicit and indirect way: different shapes are translated into different position of a same tonal target.

We decided to use a method which would ignore the specifically local features of the f0 contour (such as microprosodic minima) and at the same time avoid the implicit encoding of the global proprieties we were trying to characterize explicitly (as in the case of the two lines regression). Trying to find a compromise between these two constraints, we decided to locate the elbows at the point of maximal acceleration of the curve. Through the elaboration of an automated procedure in R (*R Development Core Team*, 2005), the L was located by inspecting the f0 second derivative, looking for sufficiently wide local maxima.

Although the L detection procedure is innovative, height and position of tone targets remain traditional measures. Besides these, we also calculated the height of the mid-point in time between L and H (*C*).[7] This allowed us to calculate an index (based on Dombrowski

[7] See Figure 4: y(C) on y-axis.

& Niebuhr, 2005) which could express the type of interpolation between the two targets in a simple and explicit way; see § 3.3.

In conclusion, for every experimental item we measured the coordinates of L, C and H in the (time, f0) plane.

3.3 Indices

We used these coordinates to calculate various indices (see Table 1), and we ran a comparison of the ability of these indices to express the contrast between the aforementioned pragmatic categories. In addition to the traditional indexes of scaling (height of L and H) and alignment (distance of L and H from both start and end of, respectively, stressed syllable onset and nucleus), we calculated a curve index, expressed as the ratio of the difference between the heights of the intermediate and the starting points, and the difference between the heights of the end and starting points of the rise.

Index	Description	Formula
sL	**L s**caling	y(L)
aLs	**L a**lignment to **s**tart of stressed vowel onset	x(L) – Os
aLe	**L a**lignment to **e**nd of stressed vowel onset	Ns – x(L)
sH	**H s**caling	y(H)
aHs	**H a**lignment to **s**tart of stressed vowel nucleus	x(H) – Ns
aHe	**H a**lignment to **e**nd of stressed vowel nucleus	Ne – x(H)
sC	**C** (intermediate point in time between L and H) **s**caling	y(C)
Ci	**C**urve **i**ndex	$\frac{y(C)-y(L)}{y(H)-y(L)}$

Table 1: Indices

4. RESULTS

The results show, for all subjects, weak or no correlations between the two pragmatic contexts (narrow focus question, QNF, and partial topic statement, SPT) and the indices usually employed in AM-based studies (alignment and scaling of tones). Two-sample Welch-Satterthwaite t-tests show that H scaling tends to be significantly different only for some speakers, while in other subjects only L scaling is significantly correlated to the two pragmatic contexts (see Figure 5).

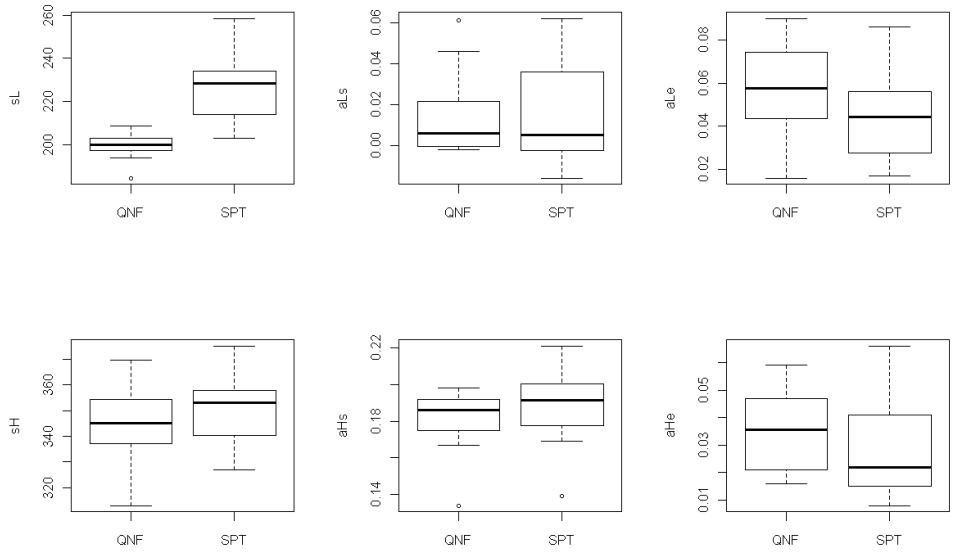


Figure 5: Box-plot for indexes sL, aLs, aLe, sH, aHs, aHe (speaker WP)

On the other hand, the curve index (and, consequently, the scaling of the midpoint in time between L and H) shows a strong correlation for all subjects with the pragmatic contexts ($p<0.001$). Moreover, considering that Ci = 0.5 would indicate a linear interpolation, we note a trend towards a convex interpolation for QNF contexts ($Ci < 0.5$), and a slight trend towards a concave interpolation for SPT contexts ($Ci > 0.5$); see Figure 6.

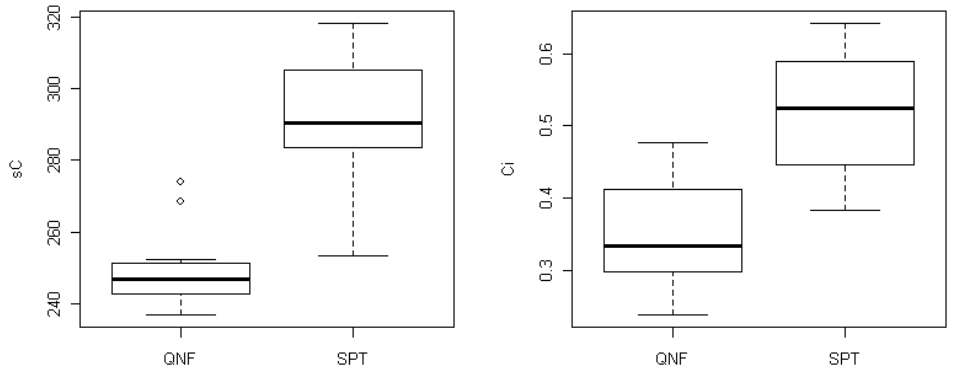


Figure 6: Box-plot for indexes sC and Ci (speaker WP)

5. DISCUSSION AND CONCLUSION

Our results confirm that two different kinds of interpolation between the two targets which compose a rising pitch accent (specifically, concave and convex) are correlated to two different pragmatic contexts (specifically, partial topic statement and narrow focus question). More generally, we can state that the analysis of the dynamic proprieties of the f0 contour allows for a better description of post-lexical meaning. If, as we mentioned at the beginning of § 2, intonational phonology investigates the relationship between suprasegmental features and post-lexical meaning, we suggest that the AM model needs to be revised in order to a give proper place to the phonological value of these dynamic properties. This claim seems to be supported by other studies, as (Petrone & D'Imperio, 2008) on NI and (Petrone & Niebuhr, 2009) on German, in which the authors examine the importance of dynamic factors outside pitch accents.

In any case, we believe that such a revision cannot be proposed before an examination is made of the perceptual relevance of the contrasts found in this production experiment. As we said in §1 (see Figures 2 and 3), even if we only take into account the f0 contour, the most striking acoustic difference between the two utterances lies in the f0 movement corresponding to the last stressed syllable (i.e., the last pitch accent). In order to correctly retrieve the pragmatic meaning of these utterances, listeners could rely mainly or exclusively on this cue. The patterns we found in production, even if robust, could prove perceptually irrelevant.

In addition, the nature of the pragmatic contrast used in this experiment is another factor that could affect the usefulness of our results. The two contexts were chosen for the acoustic features of their realizations, i.e. for the clear differences in the interpolation between the targets of the first pitch accent, which still were equally aligned and scaled. We acknowledge that from a pragmatic point of view, our two contexts are far from being prototypically contrastive. Even if the modality value of the two contexts is clearly different (question *vs* statement), both share a 'inconclusiveness' or 'openness' feature. This feature is self-evident in the question context, but it can also be retrieved in the partial topic statement. In this case, a question about the properties of a set ("How would your friends like their coffee?") is answered to by providing information on the proprieties of a subset ("As for Milena, she drinks it unsweetened…"), implicitly excluding from the predication the properties of the complement subset ("…as for the others, I couldn't say").

Should we want to use a perceptual study to evaluate the phonological importance of the phonetic opposition highlighted in the present production study, we will need to take into account these pragmatic aspects too.

ACKNOWLEDGEMENTS

I would like to thank Pauline Welby and two of the anonymous reviewers for useful comments.

6. REFERENCES

Anderson, M., Pierrehumbert, J. & Liberman, M. (1984), Synthesis by rule of English Intonation patterns, in *Proceedings of IEEE-ICASSP 84*, 2.8.1-2.8.4.

Black, A. & Hunt, A. (1996), Generating F0 contours from ToBI labels using linear regression, in *Proceedings of the 4th International Conference of Spoken Language Processing '96*, Philadelphia, October 3-6, vol 3, 1385-1388.

Boersma, P. & Weenink, D. (2009), *Praat: doing phonetics by computer*, URL http://www.praat.org/.

Büring, D. (1997), *The Meaning of Topic and Focus – the 59th Street Bridge Accent*, London: Routledge.

D'Imperio, M. & Cangemi, F. (2009), Phrasing, register level downstep and partial topic constructions in Neapolitan Italian, in *Hamburg Studies on Multilingualism,* 10 (C. Gabriel & C. Lleó, editors), Amsterdam: John Benjamins.

D'Imperio, M. (1999), Tonal structure and pitch targets in Italian focus constitutents, in *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, U.S.A., August 1-7, 1999, vol. 3, 1757–1760.

D'Imperio, M. (2000), *The role of perception in defining tonal targets and their alignment*, PhD Dissertation, The Ohio State University, U.S.A.

D'Imperio, M., Cangemi, F. & Brunetti, L. (2008), The phonetics and phonology of contrastive topic constructions in Italian, in *Third Conference on Tone and Intonation in Europe*, Lisbon, Portugal, September 15-17, 2008.

Dombrowski, E. & Niebuhr, O. (2005), Acoustic patterns and communicative functions of phrase-final F0 rises in German: activating and restricting contours, *Phonetica*, 62, 176-195.

Face, T. L. (2001), Focus and early peak alignment in Spanish intonation, *Probus*, 13, 223–246.

Hualde, J.I. (2000), Intonation in Spanish and the other Ibero-Romance languages: overview and status quaestionis, in *Romance phonology and variation* (C. Wiltshire & J. Camps, editors), Amsterdam: John Benjamins, 101–116.

Ladd, D. R. & Schepman, A. (2003), 'Sagging transitions' between high pitch accents in English: experimental evidence, *Journal of Phonetics*, 31, 81-112.

Ladd, D. R. (2008), *Intonational Phonology, 2nd edition,* Cambridge: Cambridge University Press.

Petrone, C. & D'Imperio, M. (2008), From tones to tunes: the role of f0 prenuclear region in intonation identification in Italian, in *Third Conference on Tone and Intonation in Europe*, Lisbon, Portugal, September 15-17, 2008.

Petrone, C. & Niebuhr, O. (2009), The role of the prenuclear F0 region in the identification of German questions and statements, in *Fourth Conference on Phonetics and Phonology in Iberia*, Las Palmas, Spain, June 17-19, 2009.

Pierrehumbert, J. (1980), *The phonology and phonetics of English intonation*, PhD Dissertation, Massachusetts Institute of Technology, U.S.A.

Pierrehumbert, J. (1981), Synthesizing Intonation, *Journal of the Acoustical Society of America*, 70, 985–995.

R Development Core Team (2005), *R: A language and environment for statistical computing*, http://www.R-project.org.

INTERROGATIVE E ASSERTIVE IN UN CORPUS DIALETTALE RECUPERATO (BOMARZO)

Amedeo De Dominicis
Università degli Studi della Tuscia (Viterbo)
dedomini@unitus.it

1. SOMMARIO

Il contributo si propone di analizzare un piccolo *corpus* di frasi interrogative e assertive con lo scopo di individuare un possibile descrittore fonologico e linguistico in grado di rendere conto della differenziazione tra i due tipi grammaticali.

Il *corpus* è estratto da un archivio sonoro relativo alla parlata di Bomarzo – un piccolo comune a nord di Viterbo – che fu registrato su cassette magnetiche nel 1979 e poi depositato presso l'audioteca della Provincia di Viterbo, che ne è il legittimo proprietario. Nel 2005 il materiale venne digitalizzato dal Laboratorio di Fonetica dell'Università della Tuscia,[1] sulla base di un accordo con l'ente detentore. L'archivio comprende circa 32 ore di registrazione. Si tratta di materiale eterogeneo (parti parlate si alternano a parti cantate, dialoghi a monologhi) e registrato in condizioni spesso acusticamente infelici (si trovano parti di segnale saturato, rumoroso, affetto da disturbi ambientali e da sovrapposizioni di voci). I dettagli storico-documentari sull'archivio sono illustrati in un precedente lavoro di De Dominicis & Mattana (2009).

L'interesse offerto da questo materiale è duplice. Innanzitutto, è un primo esempio di recupero di un *corpus* vocale secondo gli auspici di progetti come *Calliope* e *Covaid* (De Dominicis, 2002). Affronta e discute, quindi, problemi che potrebbero incorrere nell'analisi di qualsiasi altro *corpus* sottoposto a analogo recupero. Un secondo elemento di interesse è costituito dalla relativa profondità cronologica delle registrazioni: dati relativi a una varietà linguistica non ancora descritta e risalenti a trent'anni fa possono rappresentare un piccolo giacimento da cui attingere, sia per l'analisi sincronica che per la eventuale comparazione diacronica.

I risultati dell'indagine mostrano che la tipizzazione ToBI dell'andamento intonativo non consente di differenziare interrogative e assertive in modo consistente. Al contrario, la durata normalizzata di alcuni costituenti della gerarchia metrica appare un buon descrittore linguistico per classificare le due modalità di frase del *corpus*.

2. BASE DI ANALISI

Oggetto di analisi saranno 38 frasi interrogative e 21 frasi assertive (o, meglio, non interrogative) estratte dall'archivio in questione. Le frasi interrogative possono essere ulteriormente suddivise in aperte (14), chiuse (21), *check* (1) e polari (2).

Non è facile individuare né il numero, né – tantomeno – l'identità dei parlanti coinvolti, in quanto la documentazione associata all'archivio non consente di risalire a questi dati. Per conseguenza, su base puramente impressionistica, sembra di poter distinguere 2 parlanti di sesso maschile, anziani, e 8 di sesso femminile, per un totale di 10.

[1] Il lavoro di documentazione e riversamento digitale venne portato a termine dalla dottoressa Pamela Mattana.

Il materiale linguistico presenta un carattere spontaneo, sebbene elicitato mediante intervistatori. Sul piano sintattico, si osservano frequenti dislocazioni (più spesso a destra della frase) e focalizzazioni. I due fenomeni sono responsabili di notevoli fenomeni di allungamento di costituenti metrico-prosodici e, quindi, la loro occorrenza sarà oggetto di una attenta disamina nel successivo sviluppo dell'indagine.

3. LA VARIAZIONE DEI COSTITUENTI TONALI

Un primo tentativo di individuare un descrittore efficace della distinzione tra interrogative e non interrogative è stato condotto confrontando l'andamento intonativo dei due tipi di frase. In particolare, tale andamento è stato rappresentato mediante annotazione ToBI. Questo *standard* di etichettatura consente un confronto con studi intonativi relativi ad altre aree e varietà dell'italiano.

Studi precedenti su diverse varietà di italiano di area limitrofa hanno permesso di stabilire una tipologia di variazione diatopica relativa all'espressione dell'andamento intonativo tipico di assertive, interrogative e loro sottotipi. La tabella 1 sottostante riassume in notazione ToBI i *pitch accents* caratteristici. È presa da Sardelli (2006) ed è integrata dai dati di Giordano (2004) per Perugia, adattati da chi scrive poiché in origine non formalizzati in ToBI.

	Napoli	**Bari**	**Palermo**	**Firenze**	**Pisa**	**Lucca**	**Siena**	**Cosenza**	**Perugia**
ASS Focus Ampio	**H+L***	**H+L***	**H+L***	**H+L***	**[L+] H***	**H+L***	**H+L***		**L* L%**
ASS Focus contrastivo	**L+H***	**H*+L**	**H*+L**	**H***	**[L+] H*+L**				**H*+L L%**
INT chiuse	**L*+H**	**L+H***	**L+H***	**H***	**H+L***	**(L+H)***	**(H+L)***	**L+H***	**L* ↑L%**
Sospensive	**L***	**L***	**L***						
INT aperte						**(L+H)* L*+H**	**H* H+L***	**L+H***	**L* ↑L%**

Tabella 1: *Pitch Accents* diatopicamente caratteristici
(per Perugia si indicano anche i *Boundary Tones*)

L'analisi del nostro *corpus* ha portato a rilevare le distribuzioni caratteristiche illustrate nella Tabella 2.

Modalità	Pitch Accents (PA)	Boundary Tones (BT)
Interrogative (38=100%)	**H*+L** (15%) **!H*** (15%) H* (10%) L+H* (7%)	**L%** (55%) H% (39%)
Interrogative aperte (14=100%)	**H*+L** (28%) !H* (14%) L* L* (14%)	**L%** (57%) H% (42%)
Interrogative chiuse (21=100%)	**H*** (19%) **!H*** (19%) H*+L (9%) ↑H* +L (9%)	**H%** (47%) L% (42%)
Assertive (21=100%)	**H*** (23%) **!H*** (14%) **L*** (19%) H+L* (14%) H*+L (9%)	**L%** (76%) H% (19%)

Tabella 2: Tipi di PA nucleari e BT e relative frequenze percentuali (in grassetto le prevalenti)[2]

La distribuzione mostrata in Tabella 2 non evidenzia PA nucleari prevalenti per nessuna delle modalità di frase considerate. Per quanto riguarda le interrogative, anche volendo ritenere sommabili le percentuali di !H* e H*, si avrebbe un valore del 25%; il PA H*+L risulta prevalente nelle interrogative aperte, ma solo per il 28%; nelle interrogative chiuse la somma delle percentuali di H* e !H* equivale solo al 38% del totale; infine, per le assertive,[3] H* ricorre per il 23% e !H* per il 14% e, anche sommando i due valori, si giunge solo ad una percentuale del 37%.

Per quanto riguarda i BT prevale L% in tutte le modalità, tranne le interrogative chiuse. Ma i valori percentuali sono lungi dall'essere statisticamente significativi, dal momento che i valori possibili in questo caso sono solo due.

L'andamento intonativo delle interrogative aperte di Bomarzo sembra vagamente apparentabile a quello delle corrispondenti frasi a Siena (cfr. Tabella 1), a causa della relativa prevalenza del PA H*, ma a Bomarzo l'escursione melodica è maggiore e, comunque, tale prevalenza relativa si riscontra anche nelle interrogative chiuse di Bomarzo per le quali l'unica possibile parentela diatopica sarebbe da cercare con le corrispondenti modalità di Firenze. Parimenti, la relativa prevalenza del PA H* (e !H*) e del BT L% nelle assertive di Bomarzo potrebbe renderle comparabili a quelle (con focus contrastivo) di

[2] Per ogni modalità, si riportano solo le percentuali relative ai tipi più attestati, tralasciando i tipi di PA e BT con frequenze residuali

[3] Una classe che comprende enunciati pragmaticamente diversificati, che potremmo più correttamente definire come non-interrogative.

Perugia. Si tratta, però, di apparentamenti diatopici basati su elementi di fragile portata rappresentativa e, di regola, scarsamente affidabili.

Più in generale, ci sembra che il sistema di annotazione ToBI non riesca a fornire una tipizzazione dei contorni intonativi di Bomarzo tale da consentire un confronto tra frasi diverse. La difficoltà maggiore consiste nel fatto che si tratta di un *corpus* spontaneo, con frequentissime dislocazioni e focalizzazioni (cfr. Figure 7-11), che rendono estremamente variabile l'annotazione ToBI pur per frasi di medesima modalità. Per conseguenza, risulta molto difficile individuare un 'tipo' di PA o di BT ricorrente per ciascuna modalità analizzata.[4]

4. LA DURATA DEI COSTITUENTI DELLA GERARCHIA METRICA

Dal momento che, per quanto riguarda il *corpus* di Bomarzo, l'andamento tonale non può rappresentare un valido descrittore della distinzione tra frasi interrogative e non, abbiamo preso in esame un altro parametro fonologico. L'ipotesi da cui siamo partiti riguarda la possibilità che variazioni di durata di alcuni costituenti metrici possano essere significativamente correlate alla modalità frasale considerata.

Studi precedenti hanno già rilevato che in alcune lingue le forme interrogative non presentano un innalzamento di *pitch* in posizione finale di frase, dove invece si manifesta un sistematico incremento di durata vocalica o sillabica. In particolare, Annie Rialland (2006, 2009) ha osservato che in un rilevante gruppo di lingue di area sub-sahariana la struttura prosodica delle interrogative chiuse è 'lax' (*rilassata*), nel senso che non è espressa mediante il classico innalzamento del valore di F0, ma attraverso un allungamento vocalico finale.[5] In uno studio sulle caratteristiche prosodiche delle interrogative polari in manado malay (lingua austronesiana) e in due lingue germaniche (la varietà inglese delle isole Orkney e l'olandese), Vincent van Heuven e Ellen van Zanten (2005) rilevano l'esistenza di una diversa velocità di eloquio nelle interrogative polari rispetto alle assertive.[6]

---

[4] Un revisore anonimo lamenta la mancata sottocategorizzazione della tipologia delle frasi. In realtà, tale sottocategorizzazione è stata fatta, ma ha prodotto – come era prevedibile – una casistica talmente minuta da generare un numero enorme di tipi, di numerosità minima: quindi, privi di qualsiasi utilità statistica e metodologica.

[5] Il termine *lax question prosody* si riferisce ad un insieme di marcatori di interrogativa chiusa diffusi in area africana, che include contorno intonativo discendente, una vocale di tono basso in fine di frase, allungamento vocalico e una realizzazione *breathy* della parte terminale della frase, prodotta dalla graduale apertura della glottide; mentre, prescinde dalla tipica realizzazione delle interrogative *sì/no*: contorno intonativo ascendente o alto, tipico delle lingue indoeuropee e di molte altre aree linguistiche, tanto da essere considerato un (quasi-) universale. Il fenomeno si estende lungo la cintura sudanese africana, dall'Oceano Atlantico fino agli altopiani etiopici ed eritrei. Entro tale area la *lax prosody* si ritrova in lingue del gruppo niger-congo, specialmente nel cuore dell'area (nelle famiglie gur, kwa e kru, in cui quasi tutte le varietà hanno qualche forma di *lax prosody*). Il fenomeno è ugualmente diffuso in molte lingue del gruppo nilotico-sahariano, in particolare nelle lingue centro-sudanesi e nella famiglia delle lingua ciadiche del gruppo afro-asiatico.

[6] Il lavoro in questione analizza i correlati prosodici che segnalano la differenza tra assertive e interrogative. Il principale è l'innalzamento del *pitch* di taluni costituenti, tipico

Per l'italiano, Grazia Interlandi (2004: 272) ha confrontato le durate delle vocali nucleari (toniche) e finali (postoniche) nelle frasi assertive e interrogative nella varietà di Torino, ed ha rilevato che nelle prime la tonica e la postonica finali presentano valori assoluti inferiori rispetto alle stesse vocali della frase di modalità interrogativa; inoltre, nella frase assertiva, al contrario di ciò che avviene in quella interrogativa, la durata della tonica è solitamente maggiore rispetto a quella della postonica finale. Da questo studio sembrerebbe, quindi, che anche il parametro della durata svolga un ruolo importante nella distinzione della modalità interrogativa.

Infine, con pertinenza alla nostra discussione, si può citare un fenomeno di allungamento vocalico caratteristico di alcuni pronomi interrogativi in tamil, che viene riportato nell'antica grammatica dravidica di Robert Caldwell (1998).

Tuttavia, a nostra conoscenza, nessuno di questi precedenti studi ha preso in esame la durata di costituenti della gerarchia metrica, limitandosi, invece, a misurare le durate di elementi segmentali (vocali o sillabe).

Nel protocollo di indagine che abbiamo adottato, le 59 frasi del nostro *corpus* sono state innanzitutto analizzate metricamente: sono stati identificati i piedi metrici e, su questa base, delimitati le PW (parole fonologiche o prosodiche) e i PP (sintagmi fonologici o prosodici). In tal modo, è stato possibile ricostruire l'albero metrico di ciascuna frase. Poi, sono state misurate le durate (in ms.) di ciascun costituente metrico. Naturalmente, il valore assoluto di durata non può essere assunto come un indice di riferimento, in quanto può variare in ragione della velocità di eloquio e, quindi, del numero di segmenti pronunciati dal parlante nell'unità di tempo. Per normalizzare efficacemente i valori di durata assoluta occorrerebbe disporre di un'analisi fonologica della varietà di Bomarzo.[7] In mancanza di questa, risulta impossibile, ad esempio, stabilire se la durata di una consonante lunga o di una vocale lunga è tale fonologicamente (interpretazione monofonematica), o meno (interpretazione bifonematica).[8] Poiché la fonologia segmentale della varietà analizzata non è ancora descritta, il criterio adottato per la normalizzazione delle durate metriche è stato volutamente riduzionista: semplicemente, il valore assoluto è stato diviso per il numero dei segmenti fonetici (consonanti e vocali) associati ad un dato costituente metrico (parola o sintagma fonologico) nell'albero metrico.[9]

---

delle interrogative; quello secondario è l'aumento di velocità di eloquio, tipico sempre delle interrogative e assente nelle assertive. Le lingue analizzate sono il manado malay (una lingua austronesiana) e due lingue germaniche (l'inglese delle isole Orkney e l'olandese). In tutte e tre le lingue si rileva un incremento di velocità di eloquio nelle interrogative rispetto alle corrispondenti assertive, ma con una diversa distribuzione del fenomeno rispetto alla frase. In manado malay, la differenza sembra ristretta ai confini dei costituenti prosodici, a Orkney è ripartita uniformemente su tutta la frase, mentre in olandese si ritrova soltanto nella parte centrale della frase.

[7] Un complesso e sofisticato sistema di normalizzazione basato sul *Pairwise Variability Index* (PVI) è elaborato, ad esempio, in Grabe & Low (2002).

[8] Ad esempio, le consonanti lunghe (come la /p/ di *scappate*) sono state computate allo stesso modo di un segmento semplice (come la /p/ di *pane*). In assenza di un'analisi fonologica di questa varietà, non sembra possibile fare altrimenti.

[9] La normalizzazione delle durate metriche adottata (cioè dividendo la durata di PW per il numero di segmenti) ha lo scopo di evitare che il valore risultante sia influenzato dal numero dei segmenti. Normalizzare in base al numero dei piedi (e non dei segmenti), come

Un'ulteriore restrizione metodologica ha riguardato il tipo e la posizione dei costituenti metrici dei quali è stata misurata la durata. Per consentire un confronto tra il descrittore prosodico (annotazione ToBI dei PA e BT) e quello metrico (durata normalizzata di PW e, talvolta, PP), sono state prese in considerazione solo le PW (e – se necessario – i PP) associate a costituenti segmentali di estensione pari a quella cui sono associati i PA e i BT. In pratica, se in una frase di 10 sillabe il PA cade sulla sillaba 7 e il BT sulla 10, allora è stata considerata solo la PW i cui confini iniziano sulla sillaba 7 e finiscono sulla 10 (la PW3 in Figura 1); se su questa estensione segmentale si colloca non una, ma più PW, allora è stata misurata la durata del PP che comprende tutte le PW in questione (il PP in Figura 2). Quest'ultimo caso, in realtà, si è verificato assai di rado.

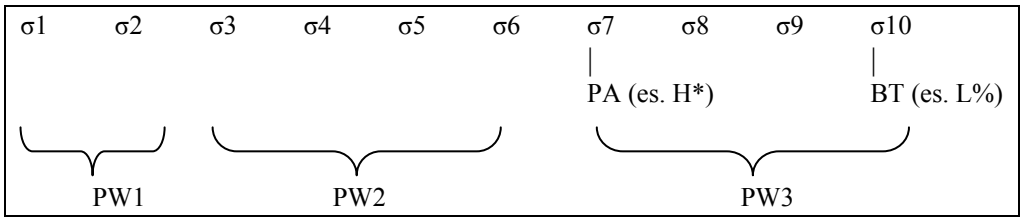


Figura 1: Simulazione di corrispondenza tra PW e PA/BT

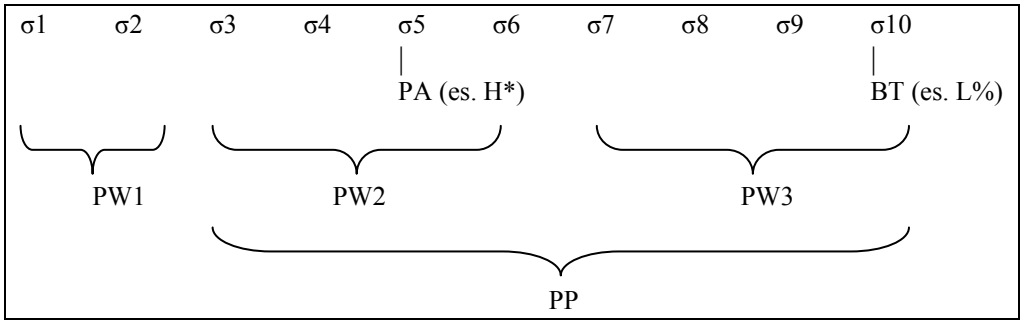


Figura 2: Simulazione di corrispondenza tra PP e PA/BT

I piedi metrici rilevati sono in maggioranza trochei (78%), alcuni giambi (17%) e altri dattili (5%). Inoltre, sono stati computati piedi degenerati (o, meglio, degeneri) e extra-metrici (Nespor, 1993; Hayes, 1995).

Per quanto riguarda le PW, la loro segmentazione è basata su criteri prevalentemente accentuali (Howell, 2004), in parziale indipendenza dalla funzione semantica/sintattica del nucleo (Giegerich, 1985; Hogg & McCully, 1987).[10]

---

suggerisce un revisore anonimo, sarebbe una procedura affidabile solo nel caso di piedi equisillabici. Diviene, invece, fuorviante qualora (come nel caso in esame) le PW siano composte da piedi con numero di sillabe variabile (piedi degeneri, binari e ternari).

[10] A seguito di un'opportuna osservazione di un revisore anonimo, segnaliamo che la definizione di parola fonologica qui adottata è solo in parte riconducibile a quella di Nespor & Vogel (1986), che pure rappresenta un classico di riferimento della fonologia prosodica. Secondo tale riferimento, il dominio di PW in italiano si identifica nella radice con tutti i

I risultati delle misurazione delle durate medie (assolute e normalizzate) realizzate mediante il protocollo suesposto sono illustrati nella Tabella 3.

Modalità	Media durata assoluta (ms.)	Media durata normalizzata
Interrogative	666	76
Interrogative aperte	626	78
Interrogative chiuse	687	74
Assertive	552	68

Tabella 3: Durate medie assolute e normalizzate delle PW

5. ALCUNI ESEMPI DI FRASI

Prima di passare all'analisi statistica dei valori di durata normalizzata delle PW nelle varie modalità, forniamo alcuni esempi tratti dal *corpus*. I dati sono analizzati con il software *Praat* ed etichettati mediante *TextGrid* a quattro-cinque livelli di *tiers*:

- primo *tier*: trascrizione IPA;
- secondo *tier*: piedi metrici e loro durate;
- terzo *tier*: PW e loro durate;
- quarto *tier*: eventuali PP e loro durate (in mancanza di PP, il contenuto del *tier* è quello del successivo);
- quinto *tier*: annotazione ToBI;
- eventuali *tiers* successivi: commenti (riguardanti il tipo di piede, la modalità di frase, ecc.).

---

suffissi e con i prefissi che terminano in consonante. Questi ultimi, a differenza di quelli uscenti in vocale, non possono costituire PW indipendenti, in quanto violerebbero una condizione di buona-formazione di struttura di parola dell'italiano (Nespor & Vogel, 1986: 129). PW contiene al massimo un accento primario, ma le parole composte consistono di due PW (Nespor & Vogel, 1986: 130).

La ragione principale di tale restrizione consiste nel fatto che il parlato qui analizzato è di tipo informale e spontaneo, mentre la base dati su cui la fonologia prosodica opera è dichiaratamente quella di un parlato né troppo rapido, né troppo lento e – soprattutto – né enfatico, né contrastivo (Nespor & Vogel, 1986: 23-24). In particolare, nel *corpus* da noi analizzato si trovano fenomeni che interferiscono con le regole di buona formazione di PW elaborate da Nespor & Vogel (1986): ad esempio, la posizione dell'accento nei composti può variare rispetto alle attese morfolessicali (es. un composto come *capostazione* potrebbe essere realizzato [ˈkapostaˌtsjone]); oppure si possono verificare cancellazioni di vocale finale di parola che impediscono la suddivisione di un composto in due PW (es. $[\text{kapo}]_{\text{PW}}$ + $[\text{statsjone}]_{\text{PW}}$ → $[\text{kapstatsjone}]_{\text{PW}}$).

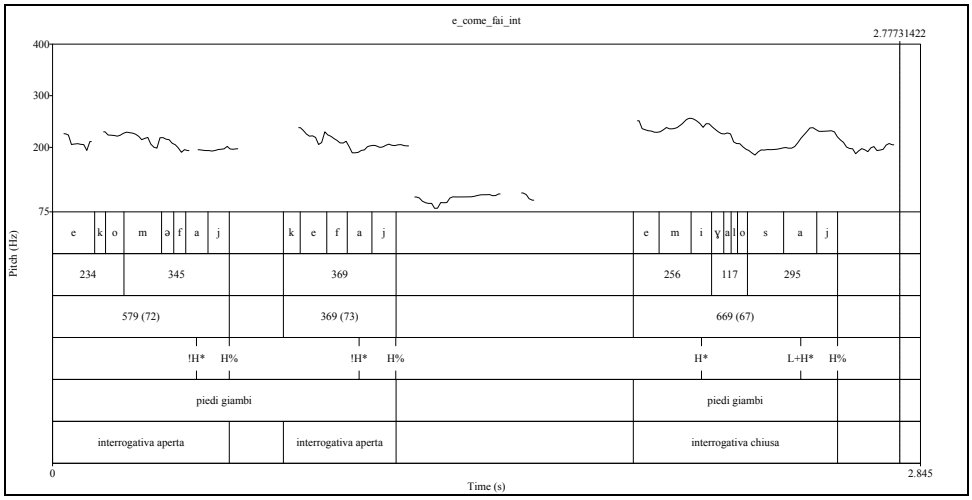


Figura 3: Esempio di piedi giambi [ekoməfajkefaj#emiɣalosaj][11] {audio 1}

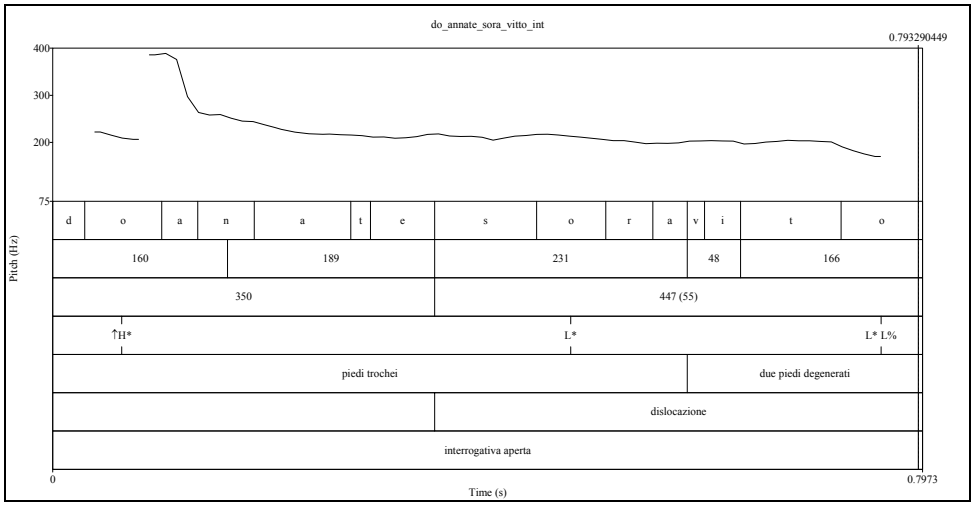


Figura 4: Esempio di piedi trochei [doanatesoravito][12] {audio 2}

[11] I piedi giambi sono sei, mentre le PW sono tre.
[12] I piedi trochei sono tre, quelli degenerati sono due, mentre le PW sono due.

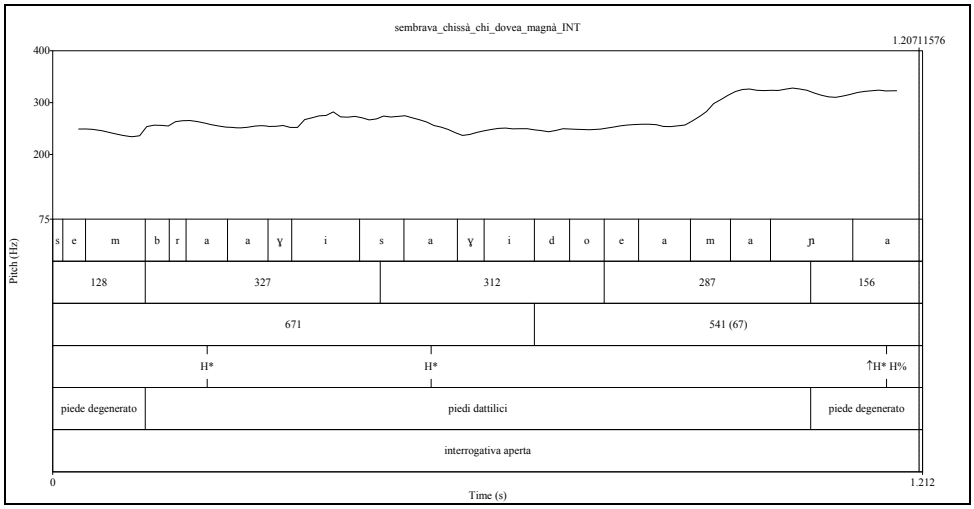


Figura 5: Esempio di piedi dattili: [sembraaɣisaɣidoeamaɲa][13] {audio 3}

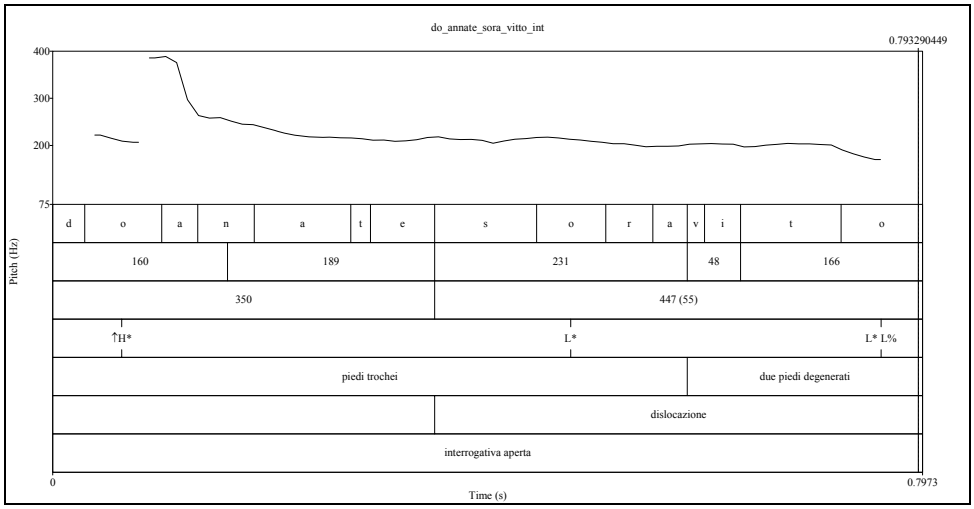


Figura 6: Esempio di interrogativa aperta: [doanatesoravito][14] {audio 4}

[13] I piedi dattilici sono tre, quelli degenerati sono due, mentre le PW sono due.
[14] I piedi trochei sono tre, quelli degenerati sono due, mentre le PW sono due.

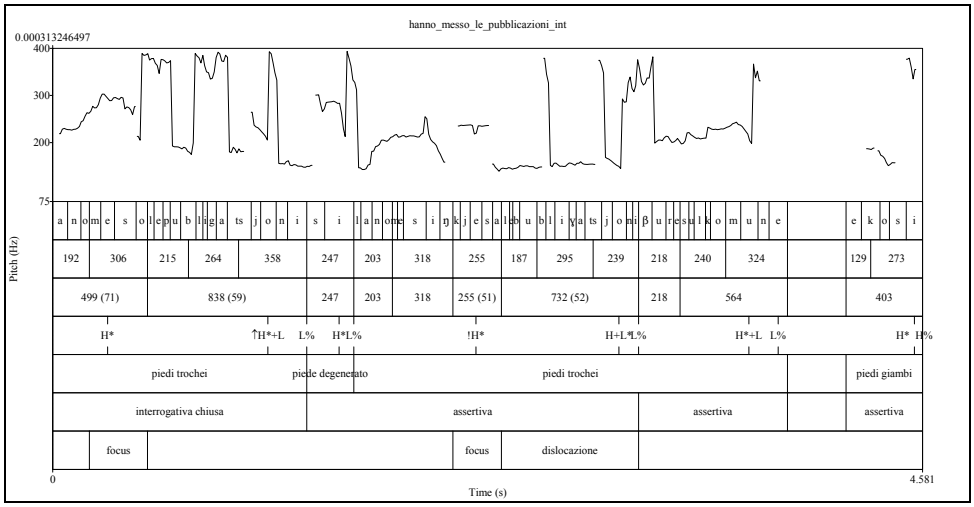


Figura 7: Esempio di interrogativa chiusa e assertive con focus [anomesolepubligatsjonisilanomesiŋkjesalebubliɣatsjoniβuresulkomuneekosi][15] {audio 5}

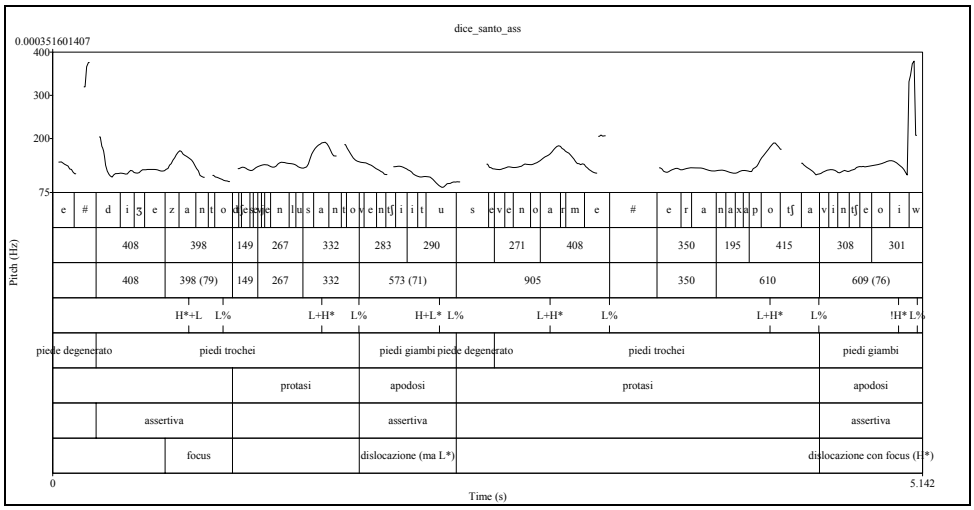


Figura 8: Esempio di assertiva con focus e dislocazione [e#diʒezantodtʃesevjenlusantoventʃiitusevenoarme#eranaxapotʃavintʃeoiw][16] {audio 6}

[15] I piedi trochei sono quattordici, uno è degenerato ed un altro giambo. Le PW sono dieci.
[16] I piedi trochei sono dieci, i giambi sono quattro, quelli degenerati sono due. Le PW sono dieci.

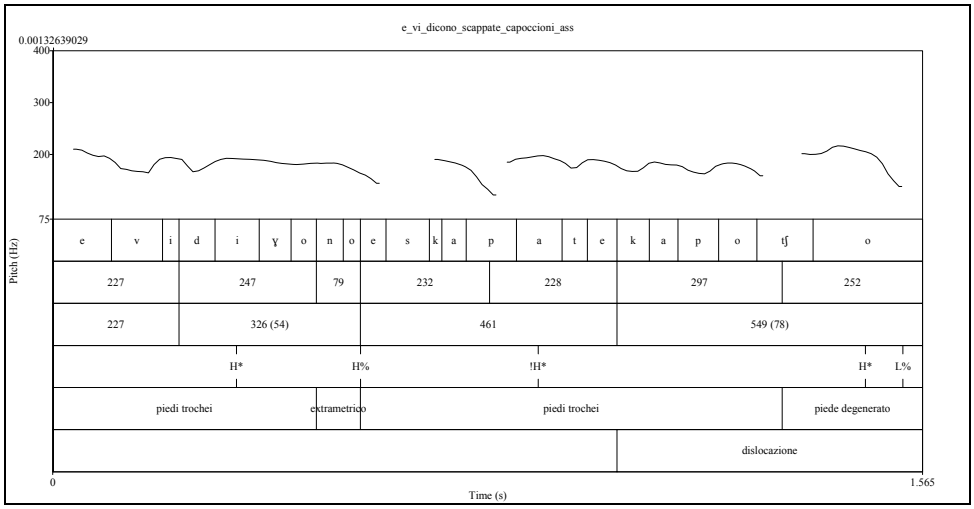


Figura 9: Esempio di assertiva con dislocazione: [evidiɣonoeskapatekapotʃo][17] {audio 7}

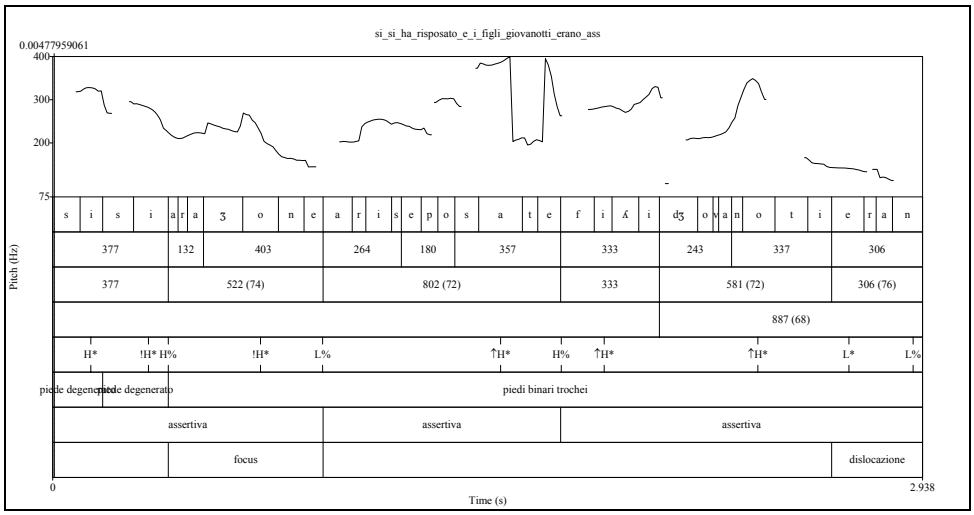


Figura 10: Esempio di assertiva con dislocazione
[sisiaraʒonearisepositefiʎidʒovanotieran][18] {audio 8}

[17] I piedi trochei sono cinque, cui si aggiungono un extrametrico ed un degenerato. Le PW sono quattro.
[18] I piedi trochei sono nove, quelli degenerati sono due. Le PW sono sei.

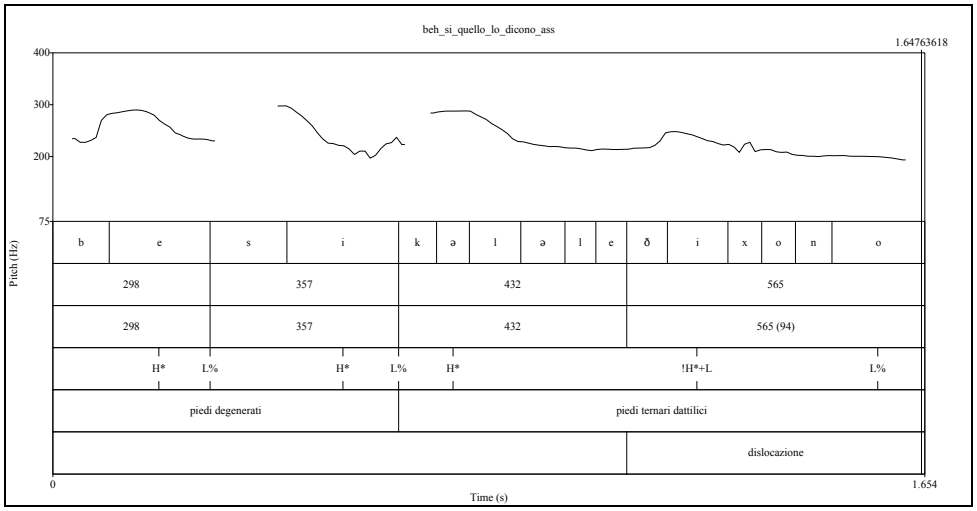


Figura 11: Esempio di assertiva con dislocazione
[besikələleðixono][19] {audio 9}

6. VERIFICA STATISTICA DELLE DURATE MEDIE NORMALIZZATE

La varianza delle durate medie normalizzate delle PW nelle interrogative (H0) e nelle assertive (H1) è riassunta nella Tabella 4:

Gruppi	***Conteggio***	***Somma***	***Media***	***Varianza***
H0: INT PW normalizzata	38	2876	75,6842	322,924
H1: ASS PW normalizzata	21	1439	68,5238	126,361

Tabella 4: Varianza delle due variabili H0 ed H1

L'analisi statistica tra le medie presentate in tabella è stata condotta con un test di Student per piccoli campioni non appaiati con varianze diverse e ad una coda. La nostra ipotesi è che non solo H0 ed H1 siano significativamente differenti, ma anche che H0 sia maggiore di H1 (le PW interrogative sono più lunghe delle PW assertive). I risultati vengono illustrati in Tabella 5.

[19] I piedi degenerati sono due, quelli dattilici sono due. Le PW sono quattro.

	INT PW normal.	*ASS PW normal.*
Media	75,68421053	68,52380952
Varianza	322,9246088	126,3619048
Deviazione standard	17,97	11,241
Osservazioni	38	21
Gradi di libertà	56	
Stat t	1,879425395	
P(T<=t) una coda	**0,032696695**	
t critico una coda	**1,672522304**	

Tabella 5: Test T a due campioni non appaiati per varianze diverse a una coda

Il valore 1,6 di t critico indica che H0 è effettivamente diversa da H1 con una percentuale del 96,8% (vedi il valore 0,032 di P(T<=t)) ed inoltre che la durata media delle PW nelle interrogative (H0) è maggiore di quella delle PW nelle assertive (H1).

Tuttavia, se osserviamo la distribuzione della varianza nei due gruppi delle interrogative e delle assertive, ci accorgiamo che, nel caso delle assertive, si rileva un'anomalia per l'eccessiva varianza di valori di durata normalizzata compresi nella fascia 75-80 e , in parte, in quella 90-95 (cfr. Figure 12 e 13).

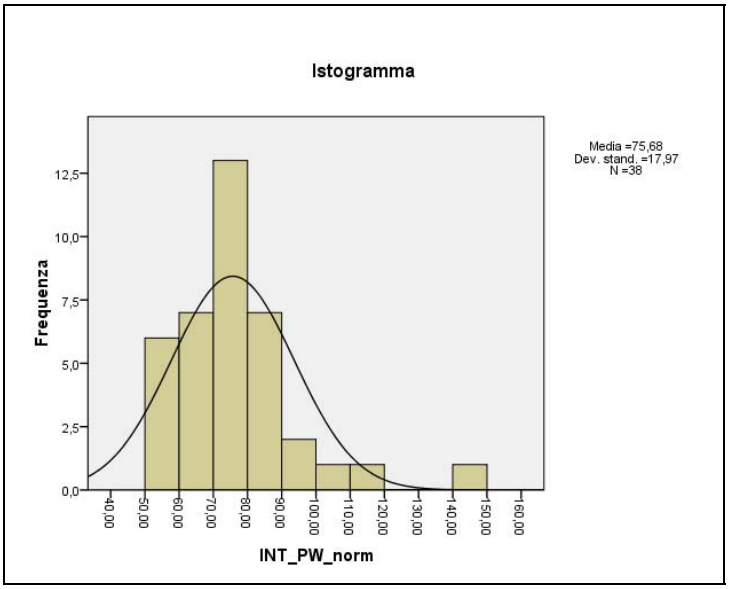


Figura 12: Distribuzione normale delle interrogative

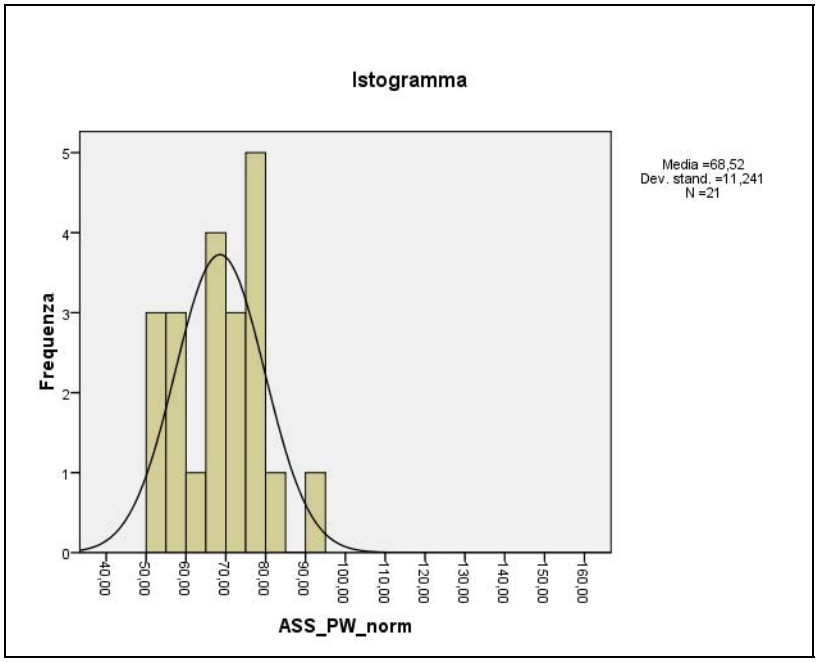


Figura 13: Distribuzione normale delle assertive

Poiché il test di Student è adatto a lavorare solo su insiemi di fenomeni supposti in distribuzione normale (cfr., ad esempio, Piccolo, 2000; Casella & Berger, 2002), i dati anomali delle assertive porterebbero ad invalidare la convalida statistica suesposta. L'alternativa, che preferiamo perseguire, è di commentare linguisticamente i casi di assertive che producono tale anomalia di distribuzione statistica normale, cioè quelli in cui le PW misurate presentano una durata normalizzata di valore compreso tra 75 e 80 e tra 90 e 95.

In effetti, si tratta di un gruppo di assertive fortemente marcate, in quanto sulle PW su cui è stata misurata la durata si collocano costituenti focalizzati o dislocati.

In dettaglio, si tratta di 4 frasi, già illustrate nelle Figure 8-11. La prima (Figura 8) presenta valori anomali sulla PW ['zanto] (durata normalizzata = 79), che è oggetto di focus, e sulla PW [vin'tʃeo'iw] (durata normalizzata = 76), che è dislocata con focus. La seconda (Figura 9) presenta valori anomali sulla PW [kapot'ʃo] (durata normalizzata = 78), che è dislocata. La terza (Figura 10) presenta valori anomali sulla PW ['eran] (durata normalizzata = 76), che è dislocata. Infine, la quarta (Figura 11) presenta valori anomali sulla PW ['ðixono] (durata normalizzata = 94), che è pure dislocata.

Un caso particolarmente istruttivo è rappresentato dalla frase assertiva in Figura 10. Qui osserviamo che le ultime due PW ([ˌdʒova'noti'eran]) hanno una durata normalizzata rispettivamente di 72 e 76. Entrambi i valori rientrano in quella fascia critica per la distribuzione normale delle assertive, di cui abbiamo appena discusso. Ma l'anomalia scompare se misuriamo la durata normalizzata su un livello più alto della gerarchia dell'albero metrico: il PP che comprende le due PW in questione ha una durata normalizzata di 68, quindi completamente al di sotto della fascia critica. In sostanza, il caso mostra che la

dislocazione sull'ultima PW ha l'effetto di renderne anomala la durata, ma la misurazione di una sequenza segmentale più ampia dei confini della dislocazione, corrispondente ad un costituente metrico di più alto grado nella gerarchia, come il PP, riconduce alla normalità i valori di durata, perché neutralizza gli effetti allunganti della dislocazione.

7. CONCLUSIONI

Abbiamo osservato che le interrogative di Bomarzo non si distinguono bene dalle non-interrogative sulla base dell'andamento intonativo: l'annotazione ToBI non riesce a fornire una tipizzazione consistente e, probabilmente, la grande variabilità sintattica e pragmatica delle frasi ostacola la comparabilità del tessuto intonativo comune alle singole modalità di frase.

Invece, sembra che la durata normalizzata di alcuni costituenti della gerarchia metrica (PW e, in parte PP) sia un buon descrittore linguistico per classificare le diverse modalità di frase del *corpus*. Sebbene studi precedenti abbiano già evidenziato una correlazione tra allungamento e interrogazione, a noi sembra che i risultati qui presentati contengano una novità. La misurazione della durata correlata alla manifestazione dell'interrogazione è computata su costituenti della gerarchia metrica e non su unità segmentali.

8. BIBLIOGRAFIA

Caldwell, R. (1998), *A Comparative Grammar of the Dravidian or South-Indian Family of Languages*, Asian Educational Services (prima edizione 1856).

Casella, G. & Berger, R.L. (2002), *Statistical inference* (2nd edition), Bemont: Duxbury.

De Dominicis, A. (2002), Co.Va.I.D. (**CO**nservazione e **VA**lorizzazione degli archivi vocali dell'**I**taliano e dei suoi **D**ialetti), *La comunicazione*, numero unico speciale a cura di Giuseppe Rinaldo e Roberto Piraino contenente gli Atti della Conferenza TIPI (Tecnologie Informatiche nella Promozione della lingua Italiana), LI, 97-98.

De Dominicis, A. & Mattana, P. (2009), Il Progetto Bomarzo, in *La Fonetica Sperimentale. Metodo e Applicazioni*, Atti del 4° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Arcavacata di Rende (CS), 3-5 dicembre 2007 (L. Romito, V. Galatà & R. Lio, editors), Torriana: EDK Editore, 405-411.

Giordano, R. (2004), *Aspetti strutturali e interrelazioni contestuali dell'intonazione dell'italiano: analisi prosodica di due dialoghi delle varietà di Roma e di Perugia*, Tesi di Dottorato non pubblicata, Università degli Studi di Perugia.

Giegerich, Heinz (1985), *Metrical phonology and phonological structure: German and English*, Cambdrige: Cambridge University Press.

Grabe, E. & Low E.L. (2002), Durational variability in speech and the rhythm class hypothesis, in *Papers in Laboratory Phonology 7* (C. Gussenhoven & N. Warner, editors), Berlin: Mouton de Gruyter, 377–401.

Hayes, B. (1995), *Metrical Stress Theory. Principles and case studies*, Chicago: University Press.

van Heuven, V.J. & van Zanten, E. (2005), Speech rate as a secondary prosodic characteristic of polarity questions in three languages, *Speech Communication*, 47, 87–99.

Hogg, R. & McCully, C.B. (1987), *Metrical Phonology: A Coursebook*, Cambridge: Cambridge University Press.

Howell, P. (2004), Comparison of two ways of defining phonological words for assessing stuttering pattern changes with age in Spanish speakers who stutter, *Journal of Multilingual Communication Disorders*, 2: 161–186.

Interlandi, G.M. (2004), *L'intonazione delle interrogative polari nell'italiano parlato a Torino: tra varietà regionale e nuova koiné*, Tesi di Dottorato non pubblicata, Università degli Studi di Pavia.

Nespor, M. (1993), *Fonologia*, Bologna: il Mulino.

Nespor, M. & Vogel, I. (1986), *Prosodic Phonology*, Berlin: Mouton de Gruyter.

Piccolo, D. (2000), *Statistica*, Bologna: il Mulino.

Rialland, A. (2006), Question prosody: an African perspective, in *Tones and Tunes: Studies in Word and Sentence Prosody* (C. Gussenhoven & T. Riad, editors), Berlin: Mouton de Gruyter, 35-62.

Rialland, A. (2009), The African lax question prosody: its realisation and geographical distribution, *Lingua*, 119, 928-949.

Sardelli, E. (2006), *Prosodiatopia: alcuni parametri acustici per il riconoscimento del parlante*, Tesi di Dottorato non pubblicata, Università degli Studi di Pisa.

BALBUZIE E COARTICOLAZIONE

Caterina Pisciotta [a], Massimiliano Marchiori [b], Claudio Zmarich [a]
[a] Istituto di Scienze e Tecnologie della Cognizione (ISTC), C.N.R., Sede di Padova;
[b] Psicologo, libero professionista
caterina.pisciotta@pd.istc.cnr.it, marchiori.massimiliano@yahoo.it, claudio.zmarich@pd.istc.cnr.it

1. SOMMARIO

La coarticolazione, intesa genericamente come l'influenza di un fono su un altro, nei soggetti balbuzienti è stata oggetto di numerosi studi: nel nostro, siamo partiti dalla considerazione che alcune ricerche sulla balbuzie suggeriscono che la coarticolazione nel parlato percettivamente fluente dei balbuzienti può essere diversa da quella riportata per i non balbuzienti – minor coarticolazione negli adulti (cfr. Robb & Blomgren, 1997), ma maggior coarticolazione dei bambini di età prescolare (cfr. Subramanian *et al.*, 2003). Esiste poi un altro filone di studi fonetici sulla balbuzie che si sofferma sulle influenze prosodiche nel linguaggio dei balbuzienti, che mostrano come essi siano effettivamente in grado di realizzare le differenze prosodiche tra parole focalizzate e non focalizzate, ma aumentano la frequenza delle disfluenze sulle parole in focus (vedi per esempio Bergmann, 1986; Marchiori *et al.*, 2005).

Questo lavoro prende le mosse dagli studi di Zmarich & Marchiori (2005) e Marchiori *et al.* (2005). In Zmarich & Marchiori (2005), le analisi su tutte le sillabe toniche (indipendentemente dalla posizione della sillaba nella parola) mostravano che i balbuzienti, rispetto ai non balbuzienti, coarticolavano di più le sillabe sotto condizione di 'focus ampio' e 'non in focus', ma di meno quelle in condizione di 'focus ristretto'. Con il presente studio, al contrario, ci siamo proposti di determinare il grado di coarticolazione anticipatoria di V su C solo sulla prima sillaba della parola, pronunciata in modo percettivamente fluente, in due contesti prosodici che favoriscono opposti livelli di coarticolazione, minimo per le sillabe non accentate e non in focus, e massimo per le sillabe accentate in focus ristretto (Zmarich, Avesani & Marchiori, 2006). I risultati evidenziano come, in condizioni massimamente critiche per il balbuziente (richiesta per un grado minimo di coarticolazione, e grado massimo di suscettibilità alle disfluenze: cfr. Marchiori *et al.,* 2005), i parlanti balbuzienti esibiscono un livello di coarticolazione significativamente maggiore rispetto a quello dei parlanti non balbuzienti. Questi risultati confermano lo studio di Subramanian *et al.* (2003), che esaminarono le transizioni di F2 nel parlato percettivamente fluente di bambini prescolari registrati subito dopo l'inizio della balbuzie.

Abbiamo inoltre analizzato le ripetizioni di sillaba nelle proposizioni disfluenti del soggetto (B) che produceva più disfluenze rispetto agli altri (54 parole disfluenti in totale). Abbiamo quindi misurato con la tecnica del *Locus of Equation* i valori di F2 al rilascio di C e al centro di V nella sillaba ripetuta immediatamente prima della sua produzione fluente (sillaba target), e abbiamo analizzato allo stesso modo anche la sillaba target e la seconda sillaba: essi appaiono piuttosto bassi, segnalando una scarsa coarticolazione. Ci sentiamo pertanto di affermare che l'alto grado di coarticolazione evidenziato dai quattro soggetti balbuzienti nella produzione fluente delle sillabe iniziali accentate in focus contrastivo possa attribuirsi a una strategia di reazione e compensazione che il soggetto balbuziente attua per tenere sotto controllo la propria balbuzie e riuscire fluente.

2. INTRODUZIONE

Con il termine 'coarticolazione' si indica l'influenza (acustica, articolatoria, percettiva) di un fono su un altro, che lungo l'asse temporale può seguirlo o precederlo. La coarticolazione viene detta perseverativa nel primo caso e anticipatoria nel secondo; cfr. il volume curato da Hardcastle & Hewlett (1999) e in particolare il capitolo – ivi contenuto – di Farnetani & Recasens (1999) per una trattazione che ne considera i molteplici aspetti.

A livello acustico, il principio della coarticolazione indica che alcune delle proprietà acustiche di un certo fono saranno presenti anche in uno o più dei foni adiacenti. Questa definizione è legata al concetto di target acustico e rivolge l'attenzione principalmente all'analisi formantica e dunque allo studio dell'andamento delle frequenze di risonanza del suono; tali indici, rilevati dall'analisi dello spettro acustico, sono particolarmente significativi in quanto permettono di risalire ai diversi atteggiamenti articolatori (luogo di costrizione principale all'interno del cavo orale secondo F2, grado di apertura orale secondo F1; cfr. Fant, 1970).

A livello percettivo, la coarticolazione, manifestandosi come compresenza delle caratteristiche acustiche proprie di un certo fono con quelle di un fono adiacente, potrebbe avere lo scopo di permettere che l'informazione relativa ad ogni dato fono possa essere mantenuta dal sistema percettivo più a lungo, favorendo un processamento in parallelo dell'informazione segmentale, e di conseguenza una trasmissione e una percezione più rapida del linguaggio.

A livello articolatorio, si ha coarticolazione quando un articolatore impegnato nella realizzazione di un fono interferisce in varia misura nella realizzazione di altri foni. L'interferenza può essere di natura fisica o acustica. Nel primo caso, in cui la coarticolazione prende il nome di adattativa, i foni coarticolati condividono uno o più articolatori necessari alla realizzazione dei loro bersagli, per es. il dorso della lingua nelle sequenze costituite da consonante velare e vocale; in questo caso, se la vocale è una vocale anteriore, come /i/, la realizzazione normale prevede che il locus articolatorio di /k/ diventi un po' più avanzato e quello di /i/ un po' meno avanzato di quelle che sarebbero le posizioni del dorso per gli stessi foni pronunciati isolatamente. Al contrario, l'interferenza può essere solo di tipo acustico (coproduzione, cfr. Fowler, 1980) quando le conseguenze acustiche del movimento degli articolatori per la produzione di foni adiacenti ad un dato fono, possono portare ad alterare la percezione di quest'ultimo rispetto a quella della sua pronuncia isolata. Questo caso viene a realizzarsi quando gli articolatori primari dei due foni sono anatomicamente e funzionalmente indipendenti, come per es. il velo, le labbra e la lingua, e il decorso temporale della loro attività si mette in relazione secondo fasi variabili di sincronizzazione che possono portare anche alla scomparsa di un certo fono, se le conseguenze acustiche del movimento articolatorio dell'altro fono si sovrappongono alle prime fino a schermarle del tutto. È questo il caso normale della produzione di sequenze formate da vocali-consonanti occlusive bilabiali-vocali, in cui durante l'occlusione bilabiale il movimento della lingua non lascia traccia acustica.

L'emergere dei fenomeni coarticolatori sembra rispondere a richieste di economia e funzionalità da parte del nostro sistema di comunicazione vocale. La coarticolazione aderisce innanzitutto ad un principio di 'economia temporale' in quanto permette al nostro unico tratto vocale, condizionato da vincoli di natura fisiologica, di produrre più segmenti fonologici, cioè più informazione, nell'unità di tempo. Inoltre il fenomeno coarticolatorio risponde anche ad un principio di 'economia da sforzo' poichè permette di diminuire lo sforzo articolatorio riducendo il contrasto fonetico e adattando un fono ai suoni che lo precedono e che lo seguono: tutti i sistemi motori, pur offrendo un insieme estremamente

ricco di possibilità per l'esecuzione di un dato compito, tendono sempre ad operare attraverso il minimo dispendio di energia.

Per quanto riguarda l'interazione tra coarticolazione e fattori prosodici (come i diversi livelli di prominenza), per la lingua inglese è stato dimostrato che le sillabe prosodicamente prominenti evidenziano una minore coarticolazione dei loro costituenti segmentali (C e V) rispetto a sillabe che sono meno prominenti (come le atone di parole che sono prive di focus, v. de Jong, Beckman & Edwards, 1993). Un maggior grado di prominenza vocalica e consonantica nelle sillabe focalizzate dovrebbe essere realizzato da labbra e lingua con movimenti tra i rispettivi target che sono più ampi e più rapidi. Questa strategia massimizza le caratteristiche di luogo di consonanti e vocali e riduce le reciproche influenze coarticolatorie. Per quanto riguarda l'italiano, Zmarich, Avesani & Marchiori (2006) hanno dimostrato come le caratteristiche fonetiche soprasegmentali influenzano la struttura prosodica del parlato ed anche la sua realizzazione segmentale: infatti le sillabe toniche sono significativamente meno coarticolate delle atone; la focalizzazione contrastiva invece, per la prima volta oggetto di studio per l'italiano, ha indotto variazioni di coarticolazione abbastanza sistematiche ma di grandezza minore e statisticamente non significative. Altri studi poi hanno messo in luce come anche la realizzazione dei confini prosodici possa essere eseguita con gradi diversi di coarticolazione che riflettono l'importanza relativa del confine all'interno della gerarchia prosodica, così che per esempio la sillaba iniziale di parola ha un grado minore di coarticolazione rispetto alle sillabe interne alla parola (cfr. Fougeron, 2001; Zmarich & Uguzzoni, 2005).

Venendo ora a discutere di balbuzie e di coarticolazione, è facile immaginare come la balbuzie, che viene definita come "un disordine nel ritmo della parola, nel quale il paziente sa con precisione ciò che vorrebbe dire, ma nello stesso tempo non è in grado di dirlo a causa di arresti, ripetizioni e/o prolungamenti di un suono che hanno carattere di involontarietà" (W.H.O, 1977; traduzione a cura di chi scrive), possa esser stata ricondotta da molti studiosi ad un'alterazione dei principi dell'economia temporale ed ergonomica che caratterizzano la coarticolazione. Da un punto di vista generale questa considerazione trae forza dalla consapevolezza che la balbuzie ha delle basi neurofisiologiche che colpiscono in particolare il sistema motorio. Numerosi studi di *Brain Imaging* condotti in questi ultimi anni sui balbuzienti mostrano anomalie nell'attivazione delle aree motorie cerebrali e del cervelletto, e tutti riportano differenze con i non balbuzienti nei controlli per la lateralizzazione cerebrale. Mentre molti risultati sono controversi, la pre-esistenza di una riduzione di densità della materia bianca nell'area motoria sinistra che presiede ai movimenti della laringe e della faccia è un fatto acclarato negli adulti (Sommer *et al.*, 2002; Sommer *et al.*, 2003) e anche nei bambini di 9 anni (Chang *et al.*, 2008; per una rassegna in italiano, vedi Patrocinio, 2008). Dal punto di vista teorico, sono ormai molti gli studiosi che spiegano la balbuzie con un tipo di controllo motorio 'debole/fragile' ma non deficitario (cfr. van Lieshout, *et al.* 2004) o deficitario *tout court* (Max, 2004), e alcuni studiosi hanno avanzato l'ipotesi che possa trattarsi di un terzo tipo di disordini motori, accanto alle famiglie ben conosciute delle disartrie e delle disprassie (Caruso & Strand, 1999; Kent, 2000; Brown *et al.*, 2005).

In questa ricerca ci siamo proposti di continuare l'analisi di Zmarich & Marchiori (2005), che avevano tentato di unificare due linee di ricerca sulla balbuzie che storicamente sono rimaste piuttosto separate, cioè quella che indaga i rapporti tra balbuzie e coarticolazione e l'altra che indaga i rapporti tra balbuzie e fattori prosodici, *in primis* l'accento. In passato queste due linee di ricerca si sono intersecate solo nell'opera di M.E. Wingate, che attribuiva la causa della balbuzie alla difficoltà di realizzare la transizione tra

l'inizio della consonante e il nucleo vocalico della sillaba, specie quando la sillaba è accentata e iniziale di parola (cfr. Wingate, 1988). In particolare:

approfondiremo l'analisi della coarticolazione della sillaba CV percettivamente fluente, per investigare cosa succede quando la sillaba da coarticolare assomma in se due fattori che le ricerche precedenti hanno messo in relazione con alte percentuali di balbuzie: la posizione iniziale di parola e il focus ristretto (per accento contrastivo);

affronteremo per la prima volta l'analisi della coarticolazione nelle sillabe disfluenti (sillabe ripetute) prima della loro realizzazione fluente, e confronteremo il grado di coarticolazione tra queste due realizzazioni della stessa sillaba, quella scorretta e quella corretta.

Prima di passare in rassegna i risultati delle diverse ricerche che hanno tentato di trovare una relazione tra balbuzie e coarticolazione, è opportuno premettere una spiegazione metodologica, senza la quale risulterebbe difficile l'interpretazione degli stessi. Esistono infatti due metodologie per misurare la coarticolazione, la *F2 ratio* (cfr. per esempio Nittrouer, Studdert-Kennedy & McGowan, 1988) e la *Locus Equation* (cfr. Sussman, Duder, Dalston & Cacciatore, 1999); bisogna sottolineare come esse a volte utilizzino gli stessi termini con significati diversi ed a volte del tutto contrapposti (vedi il termine chiave di *slope*). La F2 *ratio* e la *Locus Equation* sono state ben presentate da Farnetani (2003), della quale qui ripetiamo le parole:

"Nella metodologia F2-Ratio, gli effetti anticipatori in CV sono misurati al rilascio della consonante e sono espressi in termini del rapporto numerico tra i valori di F2 nel contesto di /i/ e F2 nel constesto di /a/. Si basa quindi sul confronto paradigmatico tra due emissioni (C/i/ vs C/a). In assenza di effetti coarticolatori il rapporto è intorno a 1, la coarticolazione aumenta quanto più il rapporto è maggiore di 1. [...] nella *locus equation* il grado di coarticolazione è rappresentato dalla regressione lineare della variabile dipendente, cioè il valore di F2 al rilascio della consonante (i.e., al V-Onset) sulla variabile dipendente, cioè il valore di F2 al V-target, secondo l'equazione: F2 Onset = K * F2 Target + C."

K, la pendenza, o coefficiente di regressione, è l'indice di coarticolazione, e varia da 0 (assenza di coarticolazione) ad 1 (massima coarticolazione). Questa procedura si basa quindi su una relazione di tipo sintagmatico tra due momenti della stessa emissione. Secondo Sussman *et al.* (1999: 770), le *locus equation* non incorrono in problemi di normalizzazione poichè "[they] are relatiolally derived and hence inherently incorporate a degree of spatial normalization".

È da notare che solo con la prima metodologia si può valutare un aspetto della coarticolazione come la velocità di transizione di F2 da C a V, poichè essa implica il riferimento alla durata della transizione (che va dal target di C, cioè il V-onset, all'inizio dello stato stabile di F2, cioè il target di V), e tale durata non può coincidere, per definizione, con il punto centrale della porzione stabile di V, misurato con la seconda metodica.

Per quanto riguarda gli adulti balbuzienti, che sono l'oggetto della presente ricerca, Stromsta & Fibiger (1981) per primi riscontrarono una riduzione della normale coarticolazione CV nelle ripetizioni dei balbuzienti, e, nell'ambito di un'analisi acustica delle vocali balbettate, Howell & Vause (1986) verificarono che la grande maggioranza delle vocali disfluenti e delle successive realizzazioni fluenti mancavano della transizione con la consonante precedente. Diversamente, Harrington (1987) trovò che nella produzione di target sillabici CV la disfluenza poteva consistere nella realizzazione della consonante foneticamente appropriata per il contesto vocalico.

Robb & Blomgren (1997), per primi, hanno analizzato acusticamente la coarticolazione linguale nel linguaggio percettivamente fluente dei balbuzienti, valutandola attraverso il tasso di variazione (escursione/durata) della transizione della seconda formante (F2) dal rilascio della consonante all'inizio della porzione stabile della vocale, che equivale alla velocità di cambiamento. In generale, i risultati hanno mostrato che la transizione CV nei balbuzienti è caratterizzata da una velocità maggiore rispetto ai nonbalbuzienti, perché secondo gli autori, la maggior velocità di transizione è correlata a cambi dimensionali nel tratto vocale più grandi o più veloci rispetto a quelli dei non balbuzienti (in sostanza, la lingua, nel passare dal target consonantico a quello vocalico, si muove di più , o a parità di tempo si muove più velocemente).

Nell'esperimento di Zmarich & Marchiori (2005), un piccolo gruppo di balbuzienti produceva in modo percettivamente fluente delle sillabe CV che erano fatte variare per contenuto segmentale, accento lessicale, focalizzazione della parola che le conteneva, posizione della sillaba nella parola e delle parole nella frase (vedi la descrizione della procedura sperimentale nel presente articolo). Per quanto riguarda la coarticolazione, le analisi su tutte le sillabe (indipendentemente dalla posizione della sillaba nella parola) mostravano che i balbuzienti, rispetto ai non balbuzienti, hanno un coefficiente di *slope* della regressione lineare che è maggiore per i valori di $F2_C$ e $F2_V$ delle sillabe atone sotto condizione di 'focus ampio' e 'non in focus' (cioè maggior coarticolazione), ma che è minore per le sillabe toniche sotto condizione di 'focus ristretto' (cioè minor coarticolazione).

Per quanto riguarda, invece, gli studi sull'influenza dei fattori prosodici sulla balbuzie, citeremo qui di seguito gli studi più significativi, rinviando per un'esposizione più approfondita a Marchiori *et al.* (2005).

Wingate (1976; 1984) diede molta importanza all'influenza delle variabili prosodiche nel manifestarsi della balbuzie, concentrando la sua attenzione soprattutto sull'accento lessicale. Descrisse pertanto la balbuzie come un difetto prosodico che si manifesta in modo intermittente sulla sillaba accentata (*stress increase*), in conseguenza dell'incapacità di realizzare la transizione tra la sillaba accentata e la successiva. Anche se in seguito cambiò idea sulla natura della transizione, identificandola nel passaggio dalla consonante alla vocale della sillaba accentata (Wingate, 1988), resta comunque uno dei primi autori ad aver considerato certe disfluenze, come le ripetizioni di sillaba, non tanto come marcatori di loci linguistici che presentano difficoltà per il balbuziente, tanto è vero che vengono prodotti spesso molte volte, quanto come la manifestazione dell'incapacità di pronunciare la sillaba successiva, la quale costituirebbe allora la vera causa della difficoltà. La sua intuizione è stata recentemente raffinata ulteriormente da Howell e collaboratori (cfr. Howell & Au-Yeung, 2002), che ne hanno fatto una pietra angolare della loro teoria EXPLAN.

Tornando a Wingate, i risultati della precedenti ricerche di Brown (1938), che avevano messo in luce la forte associazione tra episodi di balbuzie, fono iniziale di parola e classe grammaticale (parole 'contenuto'), vennero così da lui commentati: "è importante sottolineare come l'occorrenza della balbuzie nei suoni iniziali della parola è in stretta relazione con la sillaba accentata. […] Se consideriamo il fatto che la maggior parte delle parole inglesi è accentata sulla prima sillaba [e sono anche parole contenuto, nota degli aa.], non stupisce che gran parte degli episodi di balbuzie occorrano proprio su questa; ciò può essere interpretato secondo il principio che la balbuzie è dipendente dall'accento" (Wingate, 1979; trad. a cura di Marchiori *et al.*, 2005).

Weiner (1984), in apparente contrasto con l'ipotesi di Wingate che la balbuzie sia fortemente associata alla realizzazione dell'accento, sostenne che gli episodi di balbuzie erano

più legati alla produzione della sillaba iniziale in parole bisillabiche, anche quando l'accento lessicale veniva trasferito dalla prima alla seconda sillaba.

Bergmann (1986) dimostrò che i balbuzienti, nonostante incontrassero difficoltà nella realizzazione dei parametri acustici dell'accento di focus contrastivo, erano in grado di collocarlo correttamente all'interno dell'enunciato e, se balbettavano, lo facevano sulla sillaba interessata dall'accento intonativo (focus contrastivo). In un altro esperimento, che richiedeva la lettura di versi poetici caratterizzati da una successione fissa di accenti, la produzione delle sillabe accentate da parte dei balbuzienti non rispettava le caratteristiche fonetiche e fonologiche richieste dalla realizzazione normale e gli intervalli tra le sillabe accentate erano più variabili negli enunciati dei balbuzienti rispetto a quelli del gruppo di controllo. Secondo Bergmann (1986: 290), questi risultati favoriscono la conclusione che la balbuzie sia un disturbo prosodico, legato all'implementazione articolatoria dei parametri acustici che trasmettono la prominenza intonativa e accentuale.

Klouda & Cooper (1988), dimostrarono che l'accento è il locus più importante per la balbuzie: mentre i balbuzienti usano la configurazione intonativa corretta per trasmettere il focus contrastivo, gli episodi di balbuzie aumentano in modo significativo sulla stessa sillaba quando è accentata contrastivamente rispetto a quando non è accentata e si concentrano prevalentemente sulle prime parole della frase.

Prins *et al.* (1991) studiarono l'occorrenza della balbuzie nel parlato spontaneo di balbuzienti adulti in corrispondenza di sillabe sia toniche che atone e trovarono una coincidenza significativa tra episodi di balbuzie e sillabe toniche. Questa relazione, tuttavia, sembra caratterizzare solo le parole polisillabiche che seguono le prime tre della frase principale.

Jäncke *et al.* (1997)**,** attraverso due compiti sperimentali, verificarono che non vi erano differenze significative fra i balbuzienti e i non balbuzienti nel collocare correttamente l'accento in una parola, ma che invece emergevano differenze quando ai soggetti veniva chiesto, in modo per loro inaspettato, di spostare l'accento dalla seconda alla prima o dalla seconda alla terza sillaba. Questi risultati sono interpretati come supporto dell'ipotesi che la balbuzie sia un disturbo prosodico.

Bosshardt *et al.* (1997), attraverso la somministrazione di un compito in cui soggetti adulti normali e balbuzienti dovevano ripetere una frase interrogativa che veniva focalizzata alternativamente su due parole diverse, una collocata all'inizio e l'altra collocata alla fine della frase, scoprirono che le produzioni dei balbuzienti erano caratterizzabili come più monotone perché le sillabe interessate dal focus erano realizzate con escursioni inferiori di F_0 e durate più ridotte rispetto ai non balbuzienti.

Hubbard (1998), partendo dalla stretta associazione della posizione iniziale di parola con la balbuzie trovata da Weiner (1984), che però non aveva distinto l'accento lessicale da quello frasale, scoprì che i balbuzienti esibivano una quantità significativamente maggiore di balbuzie sulle sillabe iniziali di parola rispetto a quelle finali, e che le sillabe toniche non erano significativamente più balbettate di quelle atone. Inoltre i balbuzienti balbettavano maggiormente sulle sillabe iniziali di parola atone che su quelle non iniziali toniche.

Zmarich & Bernardini (2001) e Zmarich *et al.* (2001) hanno indagato l'abilità dei soggetti balbuzienti nel produrre correttamente, dal punto di vista percettivo e acustico, l'accento intonativo associato a tre tipi di focus*:* focus ampio distribuito sull'intera frase e focus ristretto contrastivo o sulla parola iniziale o sulla parola finale (vedi la descrizione della procedura sperimentale nel presente articolo). Dai risultati è emerso: i) non vi è alcuna associazione significativa degli episodi di balbuzie (cioè delle sillabe disfluenti) con l'accento intonativo correlato al focus, con l'accento lessicale e con la posizione della parola nell'enunciato (questi risultati sono in netto contrasto con le ricerche sui balbuzienti

tedeschi e statunitensi sopra citate); ii) per quanto concerne gli enunciati fluenti, l'analisi della durata delle parole (nome e verbo) e delle sillabe del nome mette in evidenza che, mediamente, quelle prodotte dai balbuzienti sono più lunghe rispetto a quelle prodotte dai non balbuzient; iii) balbuzienti realizzano il picco maggiore del contorno intonativo sul nome anche quando ad essere focalizzato è il verbo, mentre nei non balbuzienti l'accento intonativo associato alla parola in focus ristretto, sia essa nome o verbo, porta sempre la prominenza maggiore nella frase.

Marchiori *et al.* (2005), analizzando le caratteristiche delle sillabe balbettate da 6 soggetti balbuzienti adulti, hanno evidenziato un'associazione statisticamente significativa tra frequenza di balbuzie e sillaba iniziale del nome, indipendentemente da ogni altro fattore (e quindi anche dell'accento lessicale e condizione di focus). Inoltre, dalle analisi delle sillabe percettivamente fluenti di 4 soggetti balbuzienti e 4 soggetti non balbuzienti è emerso che i balbuzienti allineano il picco di F_0 entro la vocale accentata della parola in focus ristretto significativamente prima dei non balbuzienti.

Nel presente lavoro abbiamo ri-analizzato le produzioni percettivamente fluenti dei balbuzienti, registrate nell'esperimento già oggetto degli studi di Zmarich & Bernardini (2001), Zmarich *et al.* (2001), Marchiori *et al.* (2005), Zmarich & Marchiori (2005), con lo scopo di studiare la coarticolazione delle sillabe CV, pronunciate in modo percettivamente fluente, in una condizione massimamente critica per il balbuziente, cioè quando la probabilità di balbettare è massima, come nella sillaba iniziale di parola, e il livello di coarticolazione è minimo, come nella sillaba accentata in focus ristretto.

Inoltre, dato che il risultato trovato resterebbe esposto al dubbio che non si tratti di una manifestazione diretta della balbuzie, ma piuttosto della manifestazione delle strategie di compensazione e reazione del soggetto per evitare di balbettare, nel presente lavoro ci siamo proposti di analizzare anche alcune produzioni disfluenti che erano state registrate, ma non analizzate, nei succitati esperimenti di Zmarich e collaboratori. Infatti, solo se trovassimo che il livello di coarticolazione delle sillabe CV ripetute in modo disfluente prima della loro realizzazione fluente non fosse significativamente diverso da quello trovato per le sillabe in posizione critica, potremmo ipotizzare che il grado di coarticolazione riscontrato in queste ultime può essere effettivamente una manifestazione diretta della balbuzie.

3. PROCEDURA SPERIMENTALE

Dato che i soggetti e il setting sperimentale sono gli stessi già descritti in Marchiori *et al.* (2005) e Zmarich & Marchiori (2005), si rimanda a questi lavori per una esposizione più dettagliata.

3.1. Soggetti

Hanno partecipato all'esperimento sei soggetti balbuzienti adulti e quattro non balbuzienti adulti. Il test utilizzato per descrivere il livello di gravità dei soggetti balbuzienti è lo *Stuttering Severity Instrument* di Riley (1972). I soggetti balbuzienti, classificati in base alla gravità della loro patologia, erano di grado lieve (3), medio (2) e grave (1). Di questi soggetti, quelli in grado di realizzare un numero sufficiente di produzioni verbali giudicate fluenti dal punto di vista percettivo sono stati 4 (2 lievi e 2 medi). Su questi soggetti abbiamo analizzato il grado di coarticolazione della sillaba CV nelle produzioni fluenti.

Le produzioni disfluenti, invece, in questo studio sono state analizzate solo in uno dei sei soggetti balbuzienti, una locutrice di area veneta, di 25 anni di età, studentessa universitaria. Questo soggetto aveva prodotto 54 parole disfluenti. Ad una prima analisi dei dati, infatti, le produzioni disfluenti degli altri soggetti sono risultate troppo poche e/o troppo sbilanciate attraverso le varie condizioni per essere analizzabili dal punto vista statistico. Il soggetto preso in esame era stato classificato come grave.

3.2 Stimoli e procedure

I soggetti, seduti all'interno di una cabina silente, sentivano in cuffia la voce registrata di uno dei due sperimentatori che leggeva alcune brevi frasi in modalità interrogativa, al fine di indurre una risposta adeguata per il tipo di focus (vedi esempi sottostanti). L'altro sperimentatore, contemporaneamente, mostrava la frase che costituiva la risposta alla domanda, scritta su un cartoncino, attraverso il vetro della cabina: il soggetto, dopo aver udito la domanda-stimolo, rispondeva leggendo la frase sul cartoncino attraverso un microfono professionale che era collocato ad una distanza di quindici centimentri dalla bocca. La voce di ciascun soggetto era registrata a 44 kHz e 16 bit per mezzo di un registratore DAT (*Digital Audio Tape Sony DTC 1000 ES*) collegato direttamente alla cabina.

Le domande stimolo, con la risposta conseguente, erano del tipo:

a)	È dididì o dadadà che viene?	<u>dididì</u> viene
b)	Che cosa succede?	dàdada viene
c)	Viene o non viene,didìdi?	<u>viene</u> didìdi
d)	Didìdi viene o non viene?	didìdi <u>viene</u>
e)	Viene dididì o dadadà?	viene <u>dadadà</u>

Il *focus informativo* ampio o 'neutro' è sollecitato dalla domanda dell'es. (b), il *focus contrastivo* ristretto sulla parola iniziale (sottolineata nella risposte) è sollecitato dalla domanda degli esempi (a), (c), e quello sulla parola finale (sottolineata nella risposte) è sollecitato dalla domanda degli esempi (d), (e).

In questo modo, i soggetti sono stati indotti a produrre una serie di frasi dichiarative semplici Soggetto-Verbo/Verbo-Soggetto nelle quali veniva variato sistematicamente sia l'accento lessicale di una pseudo-parola trisillabica sia la struttura focale della frase di cui essa costituiva il soggetto. La sillaba tonica sia del nome che del verbo era alternativamente quindi portatrice di un accento intonativo che realizzava la prominenza frasale; dalla lette-

ratura sappiamo che tale accento è, nei non balbuzienti, qualitativamente diverso in caso di focus informativo e focus contrastivo in posizione finale di frase (Avesani, 2003).

In questo disegno sperimentale vengono minimizzate quelle condizioni (indicate da Recasens, 1999; Keating, Cho, Fougeron & Hsu, 2003) che avrebbero potuto portare a variazioni del grado di coarticolazione per variabili come :

la posizione contestuale della sillaba nella parola
(qui il nome è costituito da tre sillabe uguali);

il ruolo sintattico della parola (qui è sempre soggetto);

la velocità e l'accuratezza di elocuzione (qui sempre precisa e non veloce).

3.3. Analisi

Per quanto riguarda la prima parte dell'esperimento, ossia l'analisi della coarticolazione CV nelle sillabe delle frasi che sono state giudicate da noi come fluenti dal punto di vista percettivo, abbiamo escluso tutte quelle frasi in cui una o più sillabe erano caratterizzate da "[…] ripetizioni o prolungamenti, udibili o silenti, di brevi elementi del parlato, che hanno carattere di involontarietà: suoni, sillabe e monosillabi […]" (Wingate, 1964).

Per quanto riguarda invece l'analisi della coarticolazione CV nelle sillabe delle frasi disfluenti, abbiamo scelto di prendere in esame le ripetizioni di sillaba, poichè i prolungamenti erano troppo pochi per essere inclusi nelle analisi. Per le ripetizioni di sillaba con più di una ripetizioni, ad es. 'viene da-da-dadadà', abbiamo deciso di analizzare solo la ripetizione immediatamente precedente alla realizzazione fluente, sempre per l'esiguità del numero di questo tipo di disfluenze. Tutte le ripetizioni hanno sempre riguardato la sillaba iniziale della parola.

Abbiamo utilizzato *Praat* (Fig. 2) per calcolare lo *Spectral Slice*, basato sulla *Fast Fourier Transformation*; il metodo usato per lo studio della coarticolazione è quello delle 'equazioni di luogo' (Sussman *et al.*, 1999). È stata pertanto utilizzata la seconda delle metodologie presentate da Farnetani (2003), poichè giudicata da noi più affidabile e precisa. Le 'equazioni di luogo' si ricavano tramite le regressioni lineari dei valori di frequenza (*Hz*) della seconda formante (F2) misurati all'inizio della transizione, cioè sul primo ciclo di vibrazione glottica utile allo scopo, subito dopo il *burst* conseguente al rilascio dell'occlusione consonantica, e al centro della vocale. Secondo Sussman *et al.* (1999) tali equazioni consentono di quantificare il grado di coarticolazione anticipatoria C-V (vedi anche Zmarich & Marchiori, 2005; Zmarich *et al.*, 2005). Infatti, ogni transizione di F2 fornisce una coppia di valori frequenziali ($F2_{onset}$, $F2_{vowel}$) che, disposta nel piano cartesiano, con $F2_{onset}$ in ordinata $F2_{vowel}$ in ascissa, è in grado di individuare un punto in maniera univoca. L'insieme dei punti, relativo a una singola categoria di luogo (bilabiali, alveolari, velari) coarticolata con un'ampia gamma di vocali, si addensa lungo tutta la distribuzione dei valori ed è ben interpolato dalla retta di regressione lineare descritta dalla formula: $F2_{onset}=kF2_{vowel}+c$ (Lindblom, 1963) con k e c costanti reali che rappresentano rispettivamente la pendenza della retta di interpolazione e l'intercetta con l'asse delle ordinate (Sussman *et al.*, 1999).

Come si vede in figura 1, in una situazione ideale come quella rappresentata dal grafico in alto a sinistra, il locus consonantico, punto in cui inizia la risonanza di F2, non muta al variare del contesto vocalico e l'andamento di F2 determina una curva di transizione diretta verso i differenti valori stazionari della vocale. Avremo allora, per le diverse sillabe (CV_1, CV_2, ecc.), un insieme di coppie in cui varierà uno solo dei due valori formantici, essendo l'altro costante. Il tracciato descritto dall'equazione di luogo della retta interpolante i punti

sarà a pendenza nulla (k=0), ovvero parallela all'asse delle ascisse. Questa configurazione descrive un' assenza di coarticolazione tra la vocale e la consonante che la precede.

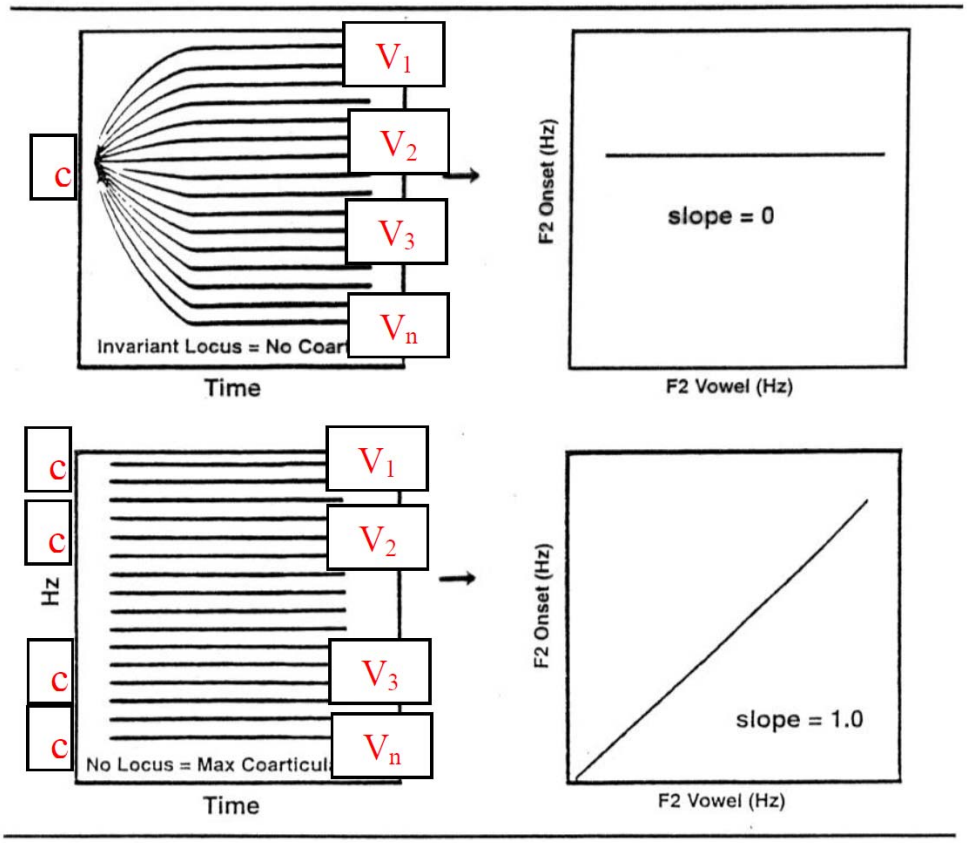


Figura 1: estremi ipotetici delle slopes (pendenze) nelle equazioni di luogo[1]

La situazione idealmente opposta è illustrata dal grafico in basso a sinistra di figura 1, in cui si verifica il massimo grado di coarticolazione anticipatoria: in questo caso non è presente alcuna transizione, poichè in ogni contesto vocalico il locus consonantico ($F2_{onset}$) sarà identico al valore dello stato stazionario della vocale ($F2_{vowel}$). Un simile insieme di coppie di valori dà luogo a punti con coordinate simmetriche che interpolati generano (per

[1] La parte della figura 1 superiore illustra le transizioni di F2 nella rappresentazione di un'assenza di coarticolazione tra la vocale e la consonante (sin.), e l'equazione di luogo con *slope* di valore 0 che risulterebbe da tale situazione (destra). La parte inferiore illustra la coarticolazione massima tra la vocale e la consonante con nessun locus consonantico fisso (sin.) e la *slope* dell'equazione di luogo risultante con valore di 1.0 (destra; modificato da Sussman *et al.*, 1999: 1082).

c=0) la retta bisettrice del piano (k=1) (Sussman et al., 1999; Zmarich e Marchiori, 2005). Questa configurazione descrive il massimo grado di coarticolazione anticipatoria.

Per ogni sillaba analizzata sono stati generati due spettrogrammi, uno a banda larga e uno a banda stretta, in modo che per ognuno degli istanti temporali sopra descritti si potessero visualizzare due sezioni spettrali. Quella dello spettro a banda stretta rappresenta le armoniche, mentre quella dello spettro a banda larga le formanti. In questo modo, confrontando le due sezioni, si è potuto individuare con chiarezza la seconda formante anche nei casi in cui l'elevato valore della fondamentale generava ambiguità e/o la formazione di artefatti formantici.

La procedura descritta è stata applicata a partire da una *routine* programmata con lo *scripting* di Praat. Quest'ultima aveva il compito di visualizzare direttamente, una volta selezionato il file della sillaba da analizzare, due *displays*, ciascuno contenente forma d'onda e sonogramma del segnale vocale, rispettivamente a banda stretta e a banda banda larga. Lo *script* utilizzato è il seguente:

Read from file... C:\pathway di allocamento del file
Edit
editor Sound nome della finestra di editor
Spectrogram settings... 0 5500 0.002 50
Formant settings... 5500 4 0.025 40 1
endeditor
Read from file... C:\pathway di allocamento del file
Edit
editor Sound nome della finestra di editor
Spectrogram settings... 0 5500 0.01 50
Formant settings... 5500 4 0.025 40 1
endeditor
Play

Da una tale visualizzazione si è poi proceduto all'estrazione delle sezioni spettrali e al loro confronto come già spiegato sopra. Le principali opzioni selezionate per l'analisi in *Praat* erano le seguenti:

a) impostazioni spettrografiche:
view range(Hz): 0-5500
window length(s): 0.002 (banda larga) / 0.1 (banda stretta)
dynamic range(dB): 50

b) impostazioni formantiche:
maximum formant(Hz): 5500
number of formants: 4
window length(s): 0.025
dynamic range(dB): 40

I due punti di misurazione della seconda formante sono (Sussman *et al.*, 1991: 1312):

F2c: valore di frequenza di F2 alla prima pulsazione glottica riconoscibile e non troppo 'rumorosa' dopo lo scoppio del rilascio dell'occlusione;

F2v: valore di frequenza a metà del nucleo vocalico, nella regione di massima ampiezza della forma d'onda e possibilmente nello stato stazionario di F2 del sonogramma.

Questo punto temporale può non essere costante a causa delle variazioni dell'andamento formantico vicino alla parte centrale del nucleo, perciò: (i) se la risonanza formantica è costituita da uno stato stazionario (determinato visivamente), allora viene selezionata la frequenza centrale dello stato stazionario; (ii) se il pattern di F2 è costituito da una discesa o da una salita con andamento diagonale, allora viene scelto visivamente di nuovo il punto centrale di questa diagonale, e (iii) se il pattern di F2 è a forma di 'U' dritta o rovesciata, allora viene preso come valore di F2 nel primo caso un minimo e nel secondo caso un massimo. Le sillabe giudicate percettivamente non conformi al target [da] o [di] non sono state considerate. Sono state inoltre escluse quelle sillabe che avrebbero potuto creare artefatti nella misurazione semplicemente a causa del metodo adottato (le equazioni di luogo), come nel caso di valori di V.O.T. positivi eccessivamente alti esibiti da qualche produzione di consonanti sorde dell'italiano. Infatti, poiché nelle equazioni di luogo i valori delle consonanti che sono messi in relazione alle vocali sono misurati sul primo ciclo glottico dopo il rilascio dell'occlusione, se questo è troppo distante dal rilascio probabilmente contiene un valore di F2 che risente dell'avvenuto riposizionamento della lingua per l'articolazione della vocale (Sussman & Modarresi, 2003). Per quanto riguarda la misurazione delle produzioni disfluenti, abbiamo misurato con la tecnica del *Locus of Equation* i valori di F2 al rilascio di C e al centro di V nella sillaba ripetuta immediatamente prima della sua produzione fluente (sillaba target), come in 'da$_1$…da$_2$dada' (dove la sillaba 1 è ripetuta e la sillaba è corretta).

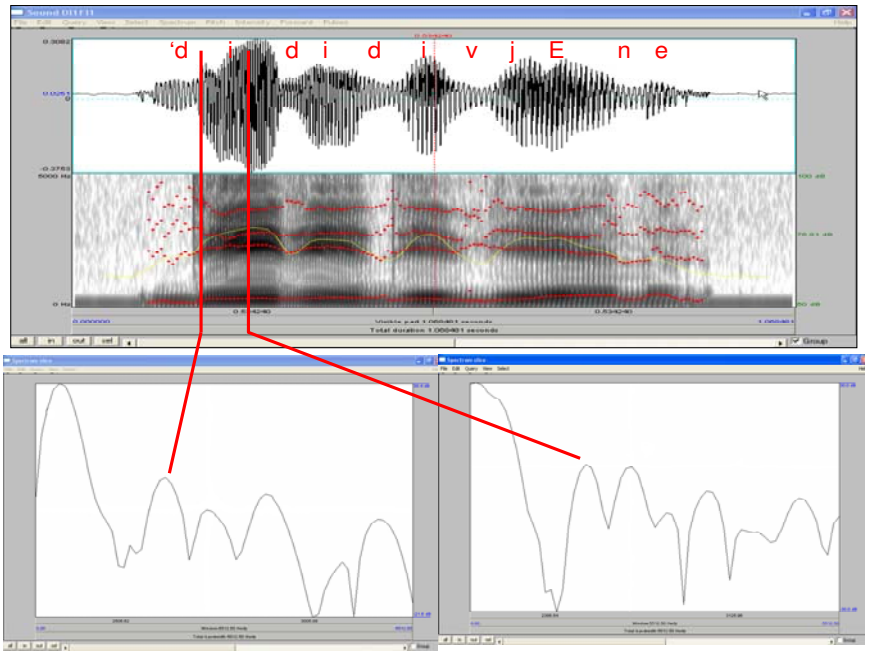


Figura 2: Finestre di analisi di PRAAT. Forma d'onda con spettrogramma (in alto) e sezione spettrale su C (in basso a sin.) e V (in basso a destra), con rilevazione dei valori di F2, sulla prima sillaba di 'dìdidi (viene)'.

4. RISULTATI

I risultati dell'analisi del grado di coarticolazione delle sillabe fluenti sono riportati in tabella 1. Essa mostra i valori individuali della slope e dell'intercetta per la prima sillaba, ritenuta la più critica per il balbuziente (grado minimo di coarticolazione, grado massimo di suscettibilità alle disfluenze). Sono state analizzate le sillabe che condividevano la proprietà di essere iniziali di parola, ma che erano poi diverse per condizione di accento lessicale e focalizzazione (atone di parole non in focus, sopra e accentate in parole in focus contrastivo, sotto). È stato calcolato, attraverso il t-test per varianze unificate (*pooled*), la significatività della differenza tra balbuzienti e non balbuzienti relativamente ai valori di *slope* e dell'intercetta, internamente a ciascuno dei due tipi di sillaba succitati.

I risultati mettono in luce una tendenza comune a balbuzienti e a non balbuzienti che era già emersa nello studio di Zmarich, Avesani & Marchiori (2006) sull'influenza di accento lessicale (di più) e accento intonativo (di meno) sulla coarticolazione dei parlanti normali, in cui le sillabe toniche erano coarticolate significativamente di meno della sillabe atone. All'interno di questa tendenza, i dati evidenziano come in condizioni di sillaba iniziale atona di parola non in focus, che *ex hypothesis* non sono critiche per i balbuzienti, i valori di *slope* dei balbuzienti sono più alti, ma non statisticamente diversi, da quelli prodotti dai non balbuzienti, mentre in condizione di sillaba iniziale tonica di parola in focus, che *ex hypothesis* è massimamente critica, i parlanti balbuzienti esibiscono un valore di *slope* significativamente maggiore rispetto a quello dei parlanti non balbuzienti (t_6=3.313, p=0.016). Ricordandoci che valori più alti e vicini all'1 per la *slope* siano associabili a un grado maggiore di coarticolazione, queste analisi ci dicono che i balbuzienti coarticolano significativamente di più dei non balbuzienti.

SILLABE NON ACCENTATE NON IN FOCUS			
SOGGETTI	STATUS	SLOPE(K)	INTERCETTA(c)
A	B	0.920	119.114
C	B	0.633	828.643
D	B	0.669	570.844
P	B	0.740	450.469
H	N	0.415	997.683
N	N	0.648	657.174
V	N	0.639	740.755
Z	N	0.671	541.113

SILLABE ACCENTATE SOTTO FOCUS CONTRASTIVO			
SOGGETTI	STATUS	SLOPE(k)	INTERCETTA(c)
A	B	0.528	901.229
C	B	0.539	1197.318
D	B	0.536	864.777
P	B	0.597	638.639
H	N	0.472	889.255
N	N	0.432	1140.963
V	N	0.462	1066.843
Z	N	0.518	850.577

Tabella 1: *Locus Equation* (valori individuali di *slope* **e** dell'intercetta solo per la prima sillaba); B = balbuzienti; N = non balbuzienti

Come già detto nell'introduzione, il risultato trovato è difficilmente interpretabile: un grado di coarticolazione significativamente maggiore nei balbuzienti rispetto ai non balbuzienti in condizioni in cui l'influsso coarticolatorio di V su C dovrebbe essere il minore possibile va inteso come una manifestazione diretta della balbuzie (una specie di restrizione di fondo che impedisce al balbuzienti di effettuare movimenti articolatori troppo estesi e/o veloci), o è da intendere invece come un'espressione delle strategie di compensazione e reazione del soggetto per evitare di balbettare e comunque tenere sotto controllo la sua balbuzie?

Forse una chiave per cercare di risolvere questo dubbio ci può essere fornita dall'analisi delle produzioni disfluenti. Infatti, se il livello di coarticolazione C-V delle sillabe ripetute in modo disfluente immediatamente prima della loro realizzazione fluente fosse significativamente maggiore di quello delle realizzazioni fluenti, potremmo ipotizzare che il grado relativamente alto di coarticolazione è un sintomo diretto di balbuzie, mentre se accadesse il contrario potrebbe essere una manifestazione delle strategie di reazione del soggetto alla sua balbuzie.

Come già esposto, abbiamo potuto purtroppo analizzare le produzioni disfluenti, in numero di 54, in un solo soggetto, che chiameremo 'B', che tra l'altro proprio a motivo della gravità della sua balbuzie, era stato escluso dall'analisi delle produzioni fluenti (cioè non faceva parte del gruppo di 4 soggetti balbuzienti confrontati con i 4 soggetti non balbuzienti a cui si riferivano le analisi delle produzioni fluenti riportate sopra). Nelle produzioni disfluenti di questo soggetto, classificate tutte come ripetizioni della sillaba iniziale (per es. da-dadada), abbiamo calcolato, per la sillaba ripetuta (da-) e per la sillaba target (dadada), l'equazione della retta di regressione (Y = k X + c) che interpola le coppie di valori F2 C /d/ e F2 V /a, i/. Va ricordato anche che il fatto di confrontare due produzioni della stessa sillaba, la prima disfluente e la seconda fluente, opera una normalizzazione intrinseca rispetto alle variabili sperimentali, pochè le due sillabe sono uguali per posizione, accento, e condizione di focalizzazione.

Come si può vedere in fig. 3, la retta non evidenzia valori di *slope* (k) significativamente diversi tra la sillaba disfluente (k= 0.541) e quella fluente (k=0.568; v. Fig. 3).

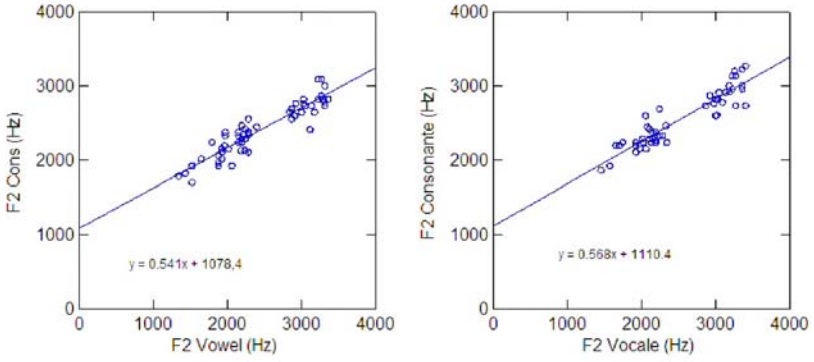


Figura 3: Diagramma F2CxF2V con l'equazione della retta di regressione (Y = k X + c) che interpola le coppie di valori F2 C /d/ e F2 V /a, i/ per le sillabe ripetute (sx) e per le sillabe target (dx) prodotte dal soggetto balbuziente B

Tuttavia, confrontando tra loro i valori assoluti, abbiamo riscontrato una differenza statisticamente significativa al t-test tra i valori di F2 C delle sillabe ripetute e i valori di F2

C delle sillabe prodotte correttamente quando la sillaba è /di/ (rispettivamente 2752 Hz *vs* 2909 Hz, p= 0,001), (dove V = circa 3100 Hz). Possiamo cautamente interpretare questi dati dicendo che quando le sillabe /di/ iniziali sono prodotte la prima volta in modo non conclusivo (cioè sono ripetizioni), sono meno coarticolate rispetto alle sillabe target, e che dunque probabilmente la lingua compie un maggior spostamento a maggior velocità passando da C a V.

Proviamo ora a immaginare che la sillaba su cui i balbuzienti incontrano difficoltà non sia la prima, ma la seconda. Infatti si potrebbe sostenere, con Wingate (1976) e Howell & Au-Yeung (2002), che quella che il balbuziente non riesce a produrre è la seconda sillaba, tanto è vero che la prima sillaba viene prodotta fin troppe volte, senza che a questa segua la seconda sillaba. Potrebbe essere dunque interessante confrontare il grado di coarticolazione intrasillabica della prima sillaba, ripetuta e fluente, e della seconda sillaba. Per fare questo confronto è però necessario comparare sillabe che sono uguali per tutte le condizioni sperimentali (di cui lo status accentuale è quello di gran lunga più importante, per il suo grande impatto sulla coarticolazione, vedi Zmarich *et al.*, 2006), eccetto ovviamente che per la posizione all'interno della parola. Per questo motivo sono state selezionate solo le prime due sillabe atone di quelle parole disfluenti che erano accentate sulla terza sillaba (10 'dadadà' e 5 'dididì'). I valori dei coefficienti ('k' e 'c') della retta di regressione per i 3 tipi di sillaba (ripetizione iniziale, prima sillaba fluente, seconda sillaba fluente) prodotti dal soggetto B (evidenziati in neretto) si possono osservare in tabella 2. Essi appaiono piuttosto bassi, segnalando una scarsa coarticolazione.

SILLABA 1			
SOGGETTI	STATUS	SLOPE(k)	INTERCETTA(c)
B (ripetiz.)	B	**0.437**	**1318.6**
A	B	0.851	254.3
C	B	0.588	898.3
D	B	0.752	458.9
P	B	0.764	388.9
B	B	**0.517**	**1218.8**
H	N	0.437	949.3
N	N	0.658	672.5
V	N	0.600	824.5
Z	N	0.714	481.5

SILLABA 2			
SOGGETTI	STATUS	SLOPE(k)	INTERCETTA(c)
A	B	0.920	109.0
C	B	0.917	164.0
D	B	0.869	232.8
P	B	0.942	59.4
B	B	**0.620**	**859.1**
H	N	0.747	361.3
N	N	0.894	168.7
V	N	0.783	440.2
Z	N	0.940	75.2

Tabella 2: *Locus Equation* (valori individuali di slope **e** dell'intercetta per la sillaba 1 (sopra) e la sillaba 2 (sotto), entrambe atone); B = balbuzienti; N = non balbuzienti

La stessa analisi è stata applicata alle prime due sillabe, atone, delle produzioni fluenti di 'dididì' e 'dadadà' dei 4 soggetti normali e 4 soggetti balbuzienti (vedi tabella 2). Il numero medio di parole prodotte da ciascun soggetto è di circa 30, equamente distribuite tra i due tipi. Il coefficiente 'k' della *slope* evidenzia una tendenza, comune ai soggetti balbuzienti e non balbuzienti, a coarticolare significativamente di meno la sillaba iniziale di parola ($t_{-3,9} = 14$; $p < 0.01$, *pooled variance*), in linea con quanto trovato, per es., da Fougeron (2001). All'interno di questa tendenza i balbuzienti coarticolano di più dei non balbuzienti, senza raggiungere la significatività statistica.

Il dato interessante che emerge però dalla tabella 2 è il confronto tra i valori di *slope* delle sillabe delle parole disfluenti prodotte dal soggetto balbuziente B e i valori di slope delle sillabe delle parole fluenti prodotte dagli altri soggetti, normali e balbuzienti. Il confronto evidenzia come le sillabe prodotte dal soggetto balbuziente B siano caratterizzate da un grado di coarticolazione anormalmente basso. Pensando che queste sillabe appartengono a parole disfluenti, viene da concludere che l'alto grado di coarticolazione evidenziato dai quattro soggetti balbuzienti nella produzione delle sillabe iniziali accentate in focus contrastivo (tab. 1), possa attribuirsi a una strategia di reazione e compensazione che il soggetto balbuziente attua per tenere sotto controllo la propria balbuzie ed essere fluente.

5. DISCUSSIONE

Il maggior risultato di questa ricerca è che il parlato percettivamente fluente dei balbuzienti mostra tendenzialmente una maggiore coarticolazione intrasillabica, così come evidenziato dai valori più alti dei coefficienti di pendenza (*slope*) della retta di regressione lineare che interpola le coppie di valori di $F2_C$ e $F2_V$, (per C=/d/ e V=/a/, /i/), rispetto ai non balbuzienti. Più in particolare, le sillabe toniche iniziali di parola in focus ristretto risultano significativamente più coarticolate. Un valore più alto della pendenza è espressione del fatto che i balbuzienti attuano, anche in una condizione critica come quella relativa a consonanti e vocali di sillabe prominenti per accento lessicale (a livello metrico), accento contrastivo (a livello intonativo) e confine prosodico (iniziali di parola), una maggior sovrapposizione gestuale tra i segmenti fonetici (comunque inferiore rispetto alle sillabe atone in focus ampio).

Già Zmarich & Marchiori (2005) e Marchiori, Zmarich, Avesani & Bernardini (2005) avevano messo in luce come la sillaba iniziale del nome sia un punto altamenlte critico per i balbuzienti. Infatti le analisi percettive mostrano che esiste un'associazione statisticamente significativa nei balbuzienti tra disfluenza e sillaba iniziale del nome, indipendentemente da ogni altro fattore.

Questi risultati sono in contrasto con lo studio di Robb & Blomgren (1997), che avevano evidenziato una minor coarticolazione nei balbuzienti adulti, ma sono in accordo con le analisi di Subramanian, Yairi & Amir (2003), che, esaminando le transizioni di F2 nelle sillabe CV percettivamente fluenti di bambini prescolari registrati subito dopo l'inizio della balbuzie, suggeriscono la presenza di eventuali deficit già allo stadio formativo del disordine. I loro soggetti erano 10 bambini balbuzienti persistenti che in seguito avrebbero cronicizzato, 10 che in seguito avrebbero avuto una remissione spontanea, e 10 normofluenti di controllo. I risultati indicano che i bambini destinati a cronicizzare avevano dimostrato cambiamenti in frequenza da C a V significativamente più ridotti rispetto al gruppo dei bambini che poi sarebbero guariti spontaneamente (e ai bambini non balbuzienti), cioè un maggior grado di coarticolazione intrasillabica. Il contrasto tra i nostri i risultati e quelli di Robb & Blomgren (1997) può essere dovuto a diversi fattori che

rendono i due studi molto diversi. I cinque balbuzienti dello studio di Robb e Blomgren erano mediamente più gravi dei balbuzienti del presente studio, e questo può significare che anche negli enunciati percettivamente fluenti facessero fatica a controllare la loro balbuzie, controllo che dovrebbe manifestarsi con una riduzione della coarticolazione. Inoltre la maggior coarticolazione esibita dai balbuzienti nel presente studio emergeva solo nel contrasto con la minima coarticolazione esibita dai parlanti normali, in un contesto sperimentale che richiedeva la massima distintività di C e V. Anche gli stimoli verbali erano diversi, poiché le sillabe in cui i balbuzienti esibivano una minor coarticolazione erano costituite da consonanti bilabiali seguite da [i], [a], o [u]. Come detto nell'introduzione, il tipo di coarticolazione esibito è definito come 'coproduzione': poiché labbra e lingua sono articolatori indipendenti, consentono la massima coarticolazione, da intendersi come coproduzione articolatoria dei gesti C e V, che risulta tanto maggiore quanto più il locutore sa utilizzare il periodo dell'occlusione bilabiale per compiere gli spostamenti linguali, in modo che all'apertura delle labbra la lingua è già in posizione per la vocale, minimizzando ogni differenza tra valori di F2 misurati all'inizio e alla fine della transizione. Questa condizione articolatoria è chiaramente differente da quella sottostante agli stimoli utilizzati nel presente studio, in cui una [d] era seguita da [i] o [a]. In questo caso C e V sono prodotte da due parti distinte (apice, dorso) e quasi indipendenti dello stesso articolatore (la lingua). Tali sillabe impongono alla lingua di muovere l'apice per articolare la consonante e il dorso per articolare la vocale (Sussman *et al.*, 1999). Questa ridotta sovrapposizione articolatoria porta a una ridotta coarticolazione intesa come co-produzione, ma lascia spazio all'emergere di influenze coarticolatorie dovute a reciproci adattamenti di dorso e apice della lingua. Quindi si può ipotizzare che la minor coarticolazione trovata da Robb & Blomgren (1997) sia dovuta non tanto a spostamenti della lingua diversi per ampiezza o velocità, quanto a spostamenti uguali che però nei balbuzienti avvengono prevalentemente dopo l'apertura delle labbra. Il problema articolatorio sembra essere di natura temporale nei soggetti di Robb & Blomgren (1997), e di natura spaziale in quelli del presente studio.

Per quanto riguarda invece le sillabe disfluenti, con questo studio ci siamo proposti di sondare se questo livello eccessivo di coarticolazione è un sintomo diretto della balbuzie, oppure una reazione reattiva, di tipo secondario ad essa. Abbiamo quindi analizzato, attraverso l'analisi acustica, le produzioni disfluenti di un soggetto balbuziente, ipotizzando che, se anche le parole disfluenti sono caratterizzate da un livello eccessivo di coarticolazione, è facile concludere che essa è allora un sintomo diretto della balbuzie, che si manifesta anche nelle produzioni percettivamente fluenti. Il confronto evidenzia come le sillabe prodotte dal soggetto balbuziente B siano caratterizzate da un grado di coarticolazione anormalmente basso. Pensando che queste sillabe appartengono a parole disfluenti, viene da concludere che il grado di coarticolazione, significativamente più alto di quello dei non balbuzienti, evidenziato dai quattro balbuzienti nella produzione delle sillabe iniziali accentate in focus contrastivo, possa attribuirsi a una strategia di reazione e compensazione che i soggetti di questo studio attuano per tenere sotto controllo la propria balbuzie.

Per concludere vogliamo sottolineare il fatto che questi risultati, per poter assurgere a valore generale, necessitano di trovare conferma in un numero di soggetti più ampio di quello del presente studio; uno studio futuro inoltre dovrebbe presentare il confronto tra le produzioni fluenti e disfluenti all'interno degli stessi soggetti, cosa che qui non è stata possibile.

6. BIBLIOGRAFIA

Avesani, C. (2003), La prosodia del focus contrastivo. Un accento particolare?, in *La Coarticolazione* (G. Marotta & N. Nocchi, editors), Atti delle XIII Giornate del Gruppo di Fonetica Sperimentale, Pisa, 28-30 Novembre 2002, 157-167.

Bergmann, G. (1986), Studies in stuttering as a prosodic disturbance, *Journal of Speech and Hearing Research*, 47, 778-782.

Bosshardt, H.G., Sappok, C., Knipschild, M. & Hölscher, C. (1997), Spontaneous imitation of fundamental frequency and speech rate by nonstutterers and stutterers, *Journal of Psycholinguistic Research,* 26, 425–448.

Brown, S.F. (1938), Stuttering with relation to word accent and word position, *Journal of Abnormal Social Psychology*, 33, 112-120.

Brown, S., Ingham, R.J., Ingham, J.C., Laird, A.R. & Fox, P.T., (2005), Stuttered and fluent speech production: an Ale meta analysis of functional neuroimmaging studies, *Human Brain Mapping*, 25, 105-117.

Caruso, A.J & Strand, E.A. (1999), Motor speech disorders in children: definitions, background, and a theoretical framework, in *Clinical Management of Motor Speech Disorders in Children* (A. Caruso & E. Strand , editors), New York: Thieme, 1-27.

Chang, S.E., Erickson, K, Ambrose, N., Hasegawa-Johnson, M. & Ludlow, C. (2008), Brain anatomy differences in childhood stuttering, *NeuroImage*, 39, 1333-1344.

De Jong, K., Beckman, M.E., Edwards, J. (1993), The interplay between prosodic structure and coarticulation, *Language and Speech*, 36, 197-212.

Fant, G. (1970), *Acoustic theory of speech production*, The Hague: Mouton.

Farnetani, E. (2003), The supralaryngeal articulation of prominence in Italian vowels, in *Voce, Canto, Parlato. Studi in onore di Franco Ferrero* (P. Cosi, E. Magno Caldognetto & A. Zamboni, editors), Padova: Unipress, 149-155.

Farnetani, E. & Recasens, D. (1999), Coarticulation Models in Recent Speech Production theories, in *Coarticulation: Theory Data and Techniques* (W.J. Hardcastle & N. Hewlett, editors), Cambridge (UK): Cambridge University Press, 31-65.

Fougeron, C. (2001), Articulatory properties of initial segments in several prosodic constituents in French, *Journal of Phonetics*, 29, 109-135.

Fowler, C.A. (1980), Coarticulation and theories of extrinsic timing, *Journal of Phonetics*, 8, 113-133.

Hardcastle, W. J. & Hewlett, N. (1999), *Coarticulation. Theory, Data and Techniques*, Cambridge (UK): Cambridge University Press.

Harrington, J. (1987), Coarticulation and stuttering: an acoustic and electropalatographic study, in *Speech Motor Dynamics in Stuttering* (H.F.M. Peters & W. Hulstijn, editors), Wien-New York: Springer-Verlag, 381-392.

Howell, P. & Vause, L. (1986), Acoustic analysis and perception of vowels in stuttered speech, *Journal of the Acoustical Society of America,* 79, 1571-1579.

Howell, P. & Au-Yeung, J. (2002), The EXPLAN theory of fluency control and the diagnosis of stuttering, in *Pathology and therapy of speech disorders* (E. Fava, editor), Amsterdam: John Benjamins, 75-94.

Hubbard, C.P. (1998), Stuttering, stressed syllables, and word onsets, *Journal of Speech, Language and Hearing Research*, 41, 802-808.

Jäncke, L., Bauer, A. & Kalveram, K.T. (1997), Prosodic disturbances in stuttering adults, in *Speech production: Motor control, brain research and fluency disorders* (W. Hulstijn, H.F.M. Peters & P.H.H.M. van Lieshout, editors), Amsterdam: Excerpta Medica, 479-486.

Keating, P.A., Cho, T., Fougeron, C., & Hsu, C.S. (2003), Domain-initial articulatory strengthening in four languages, in *Phonetic Interpretation. Papers in Laboratory Phonology VI* (J. Local, R. Ogden, R. Temple, editors), Cambridge (UK): Cambridge University Press, 145-163.

Kent, R.D. (2000), Research on Speech Motor Control and its disorders: a review and prospective, *Journal of Communication Disorders*, 33, 391-428.

Klouda, G.V. & Cooper, W.E. (1988), Contrastive stress, intonation, and stuttering frequency, *Language and Speech,* 31, 3–20.

Lindblom, B. (1963), On vowel reduction, *The royal institute of technology, speech transmission laboratory*, Stockholm, Sweden, Report No 29.

Marchiori, M., Zmarich, C., Avesani, C. & Bernardini, S. (2005), Focus e prosodia nelle produzioni verbali dei balbuzienti, in *Misura dei parametri. Aspetti tecnologici ed implicazioni nei modelli linguistici.* Atti del 1° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Padova, 2-4 dicembre (P. Cosi, editor), Brescia: EDK Editore, 251-286.

Max, L. (2004), Stuttering and internal models for sensorimotor control: A theorerical perspective to generate testable hypotheses, in *Speech motor control in normal and disordered speech* (B. Maassen, R.D. Kent, H.F.M. Peters, P.H.H.M. van Lieshout, W. Hulstijn, editors), Oxford (UK): Oxford University Press, 357-387.

Nittrouer, S., Studdert-Kennedy, M. & McGowan, R. (1988), The Emergence of Phonetic Segments: Evidence from the Spectral Structure of Fricative-Vowel Syllables Spoken by Children and Adults, *Journal of Speech and Hearing Research*, 32, 120-132.

Patrocinio, D. (2008), Brain imaging e balbuzie, *Acta Phoniatrica Latina,* 30, 170-201.

Prins, D., Hubbard, C.P. & Krause, M. (1991), Syllabic stress and the occurrence of stuttering, *Journal of Speech Hearing Research*, 37, 1011-1016.

Recasens, D. (1999), Lingual Coarticulation, in *Coarticulation. Theory, Data and Techniques* (W.J. Hardcastle & N. Hewlett, editors), Cambridge (UK): Cambridge University Press, 80-104.

Riley, G. (1972), A stuttering severity instrument for children and adults, *Journal of Speech and Hearing Disorders*, 37, 314-322.

Robb, M. & Blomgren, M. (1997), Analysis of F2 transitions in the speech of stutterers and nonstutterers, *Journal of Fluency Disorders,* 22, 1-16.

Sommer, M., Koch, M.A., Paulus, W., Weiller, C. & Buchel, C. (2002), Disconnection of speech-relevant brain areas in persistent developmental stuttering, *The Lancet*, 360, 380-383.

Sommer, M., Wischer, S., Tergau, F. & Paulus, W. (2003), Normal introcortical excitability in developmental stuttering, *Movement Disorders*, 18, 826-830.

Stromsta, C & Fibiger, S. (1981), Physiological correlates of the core behaviour of stuttering, in *Proceedings of the 18th Congress* of the *International Association of Logopedics and Phoniatrics,* Washington, D.C.: American Speech-Language-Hearing Association.

Subramanian, A., Yairi, E. & Amir, O. (2003), Second formant transition in fluent speech of persistent and recovered preschool children who stutter, *Journal of Communication Disorder*, 36, 59-75.

Sussman, H.M., McCaffrey, H. & Matthews, S. (1991), An investigation of locus equations as a source of relational invariance for stop place categorization*, Journal of the acoustical society of America*, 90, 1309-1325.

Sussman, H. M., Duder, C., Dalston, E. & Cacciatore, A. (1999), An acoustic analysis of the developmental of CV coarticulation: A case study, *Journal of Speech, Language and Hearing Research,* 42, 1080-1096.

Sussman, H. M. & Modarresi, G. (2003), The stability of Locus Equation encoding of stop place, in *Proceedings of 15th International Congress of Phonetic Sciences*, Barcelona, 1931-1934.

Van Lieshout, P.H.H.M., Hulstijn, W. & Peters, H.F.M. (2004), Searching for the weak link in the speech production chain of people who stutter: A motor skill approach, in *Speech motor control in normal and disordered speech* (B. Maassen, R.D. Kent, H.F.M. Peters, P.H.H.M. van Lieshout & W. Hulstijn, editors), Oxford (UK): Oxford University Press, 313-355.

Weiner, A. (1984), Stuttering and syllabic stress, *Journal of. Fluency Disorders*, 9, 301-305.

Wingate, M.E. (1976), *Stuttering: Theory and treatment*, New York: Irvington.

Wingate, M.E. (1979), The loci of stuttering: Grammar or prosody?, *Journal of Comunication Disorders,* 12, 283-290.

Wingate, M.E. (1984), Stutter events and linguistic stress, *Journal of Fluency Disorders*, 9, 295-300.

Wingate, M.E. (1988), *The structure of stuttering (a psycholinguistic analysis)*, New York-Wien: Springer Verlag.

World Health Organization (1977), *Manual of the International statistical classification of deseases, injuries, and causes of death*, 1, Geneva: World Health Organization.

Zmarich, C., Avesani, C. & Bernardini, S. (2001), La balbuzie come disturbo prosodico. Dati sperimentali su soggetti italiani, in *Multimodalità e Multimedialità nella Comunicazione* (E. Magno Caldognetto & P. Cosi, editors), Atti delle XI Giornate di Studio del Gruppo di Fonetica Sperimentale, Padova, 29-30 novembre 2000, 157-164.

Zmarich, C., Avesani, C. & Marchiori, M. (2007), Coarticolazione e Accentazione, in *Scienze Vocali e del Linguaggio – Metodologie di Valutazione e risorse Linguistiche* (V. Giordani, V. Bruseghini & P. Cosi, editors), Atti del 3° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Trento, 29 novembre – 1° dicembre 2006, Torriana (RN): EDK Editore, 5-15.

Zmarich, C. & Bernardini, S. (2001), Contrastive stress in Italian stutterers, in *Proceedings of the Third World Congress on Fluency Disorders*, August 7-11, 2000, Nyborg (DK) (H.G. Bosshardt, S. Yaruss & H.F.M. Peters, editors), Nijmegen (NL): Nijmegen University Press, 256-260.

Zmarich, C. & Marchiori, M. (2004), L'influenza del focus contrastivo sulla coarticolazione anticipatoria di sillabe 'CV' prodotte fluentemente da balbuzienti e non balbuzienti, in *Misura dei parametri. Aspetti tecnologici ed implicazioni nei modelli linguistici* (P. Cosi, editor), Atti del 1° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Padova, 2-4 dicembre 2004, Brescia: EDK Editore, 231-250.

Zmarich, C. & Uguzzoni, A. (2005), Confini di sillaba, confini di parola e lunghezza fonologica delle vocali in area frignanese: analisi cinematica dei gesti labiali, in *Analisi Prosodica. Teorie, modelli e sistemi di annotazione* (R. Savy & C. Crocco, editors), Atti del 2° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Salerno, 30 novembre - 2 dicembre 2005, Brescia: EDK Editore, 612-631.

CANTO

OSSERVAZIONI PRELIMINARI SUGLI ASSETTI INTERVALLARI NEL CANTO *A MUTETUS* DELLA SARDEGNA MERIDIONALE

Paolo Bravi
Conservatorio di Musica "G. P. da Palestrina", Cagliari
pa.bravi@tiscali.it

1. SOMMARIO

La ricerca sui caratteri delle scale musicali, oltre ad essere stata uno degli aspetti più rilevanti dell'analisi in ambito etnomusicologico fin dalle origini della disciplina (Ellis, 1885; Hornböstel, 1913), è un elemento apparentemente ineludibile nella definizione dei caratteri di un "sistema musicale". È stato frequentemente osservato, infatti, che le strutture intervallari relative a stili di canto non assimilabili ai generi della musica 'classica' (o di sua derivazione) diffusi in ambito extraeuropeo e folklorico hanno caratteristiche diverse rispetto al sistema euroculto, in particolare per quanto riguarda l'ampiezza degli intervalli, raramente assimilabile a quelli della scala temperata che da circa tre secoli costituisce il sistema di riferimento in ambito occidentale.

Il presente lavoro prende in esame uno stile di canto utilizzato dai poeti improvvisatori dell'area meridionale della Sardegna (*cantadoris*) per l'esecuzione di componimenti nella forma metrica del *mutetu*. La metodologia adottata è quella inaugurata, agli inizi degli anni Settanta, da Tjernlund, Sundberg e Fransson (Tjernlund *et al*., 1972). Il presupposto di base di tale metodo è che le frequenze prevalenti in un'esecuzione siano il correlato acustico della scala in uso, per cui l'analisi della distribuzione del pitch può offrire indicazioni essenziali sugli assetti intervallari sui quali si basa la pratica musicale. Nello stile di canto esaminato essi appaiono assimilabili solo in misura limitata a quelli definiti dal sistema temperato.

2. INTRODUZIONE

Uno degli elementi che usualmente è considerato, seppure con cautela e senza pretese di universalità, un aspetto potenzialmente discriminante fra l'intonazione nel parlato e nel canto è la stabilizzazione delle altezze, cioè il fatto che i movimenti melodici delle voci nel canto si articolino principalmente attorno ad altezze definite e ricorrenti nel corso delle esecuzioni.[1] Le curve melodiche del parlato, a prescindere dalle altezze assolute e dall'articolazione ritmica, manifestano in genere una "libertà" che di norma non si osserva nel canto. In coerenza con questo assunto, mentre gli strumenti utilizzati per l'annotazione prosodica danno una descrizione/interpretazione essenziale dei movimenti melodici,[2] le

[1] Francesco Giannattasio, pur rilevando "l'estrema varietà di forme situabili in un *continuum* fra parlato e cantato", individua nel "livellamento" uno dei procedimenti attraverso cui si realizza l'"alterazione fonica della parola ai fini di una sua particolare formalizzazione espressiva" (Giannattasio, 2005: 1013-1014). Cfr (List, 1963; Sorce Keller, 1990).

[2] In ambito linguistico, i metodi di rappresentazione e di interpretazione dell'intonazione variano considerevolmente in base al tipo di approccio, in particolare in relazione all'orientamento fonetico/fonologico dell'analisi. Tra i metodi di trascrizione più diffusi e noti attualmente ricordo qui il sistema ToBI, che mira ad una rappresentazione fonologica

trascrizioni del canto – e le trascrizioni musicali in genere, a meno che non si tratti di strumenti ad intonazione indeterminata – indicano solitamente con precisione le altezze utilizzate.[3] La strutturazione di tali altezze sulla base di rapporti sostanzialmente stabili e definiti è alla base della definizione dei *gamut* e rappresenta un presupposto essenziale dei concetti di *modo* e di *scala*, ampiamente utilizzati nell'ambito dell'analisi dei sistemi musicali.[4]

Uno dei metodi di indagine sull'organizzazione intervallare di un'esecuzione musicale è quello che si basa sull'analisi della distribuzione di frequenza dei valori di F0 (cfr Tjernlund *et al.*, 1972; Bel, 1998; Van der Meer, 2000). Questo metodo permette di osservare la presenza di aree di relativa stabilità nel canto. Una distribuzione dei valori di F0 concentrata attorno a determinate aree indica infatti che il cantore usa relativamente spesso quelle frequenze e/o vi si sofferma relativamente a lungo. Almeno in via di ipotesi, dunque, si possono considerare tali frequenze come il correlato acustico dei gradi di una scala e si può tentare di definire le relazioni fra tali gradi attraverso la misura dell'ampiezza degli intervalli.[5]

L'analisi che qui presento riguarda uno stile di canto diffuso nella Sardegna meridionale, quello del canto *a mutetus*. Si tratta di uno stile di canto legato alla pratica della poesia estemporanea, una pratica che ha come momento di vertice la performance della *cantada* (gara poetica), cui partecipano poeti improvvisatori (*cantadoris*) semi-professionisti utilizzando la forma metrica del *mutetu* (anche definito, per esigenze di

---

del contorno intonativo, descritto come una sequenza di toni alti (H) e toni bassi (L) (Pierrehumbert & Beckman, 1988), e il sistema Momel-INTSINT che si basa su una modellizzazione del contorno intonativo e una rappresentazione fonologica "di superficie" realizzata attraverso un codice costituito da otto simboli – T (Top), M (Mid), B (Bottom), H (Higher), L (Lower), S (Same), U (Upstepped), D (Downstepped) – (Hirst & Di Cristo, 1998).

[3] Per un'introduzione all'argomento, vd. (Ferrari, 1999).

[4] Il concetto di *gamut* si riferisce a "uno schema scalare "neutro" ", in cui "[t]utti i gradi della scala e gli intervalli individuati in un dato corpus musicale sono riportati su un pentagramma, spesso nell'ordine in cui essi compaiono (ad esempio in un dato canto), senza inferire da essi alcuna gerarchia tra i suoni" (Giuriati 1991: 107). Per una discussione comparativa e critica dei concetti di *modo* e *scala*, vd. (Powers, 1980; Powers, 1992; Nattiez, 1981).

[5] La terminologia qui adottata è di uso comune nell'ambito dell'analisi musicologica ed etnomusicologica. Per i non addetti ai lavori, indico in modo sintetico e deliberatamente aproblematico che per 'scala' si intende in termini generali il "repertorio di altezze", collocate in sequenza, su cui si basa – di norma – una pratica musicale; per 'gradi della scala' si intendono "le varie note di una scala, conteggiate dal basso verso l'alto"; per 'intervallo' si intende la "distanza fra due altezze"; le definizioni sono riprese da Baroni, (2004: 33, 244 e 36 rispettivamente), cui rimando per un'efficace e sintetica introduzione non tecnica all'analisi musicale. Preciso che nel caso del canto *a mutetus* non esiste una "etnoteoria" (Cardona, 1985; Baily, 2005) sistematica che permetta di inquadrare i dati rilevabili all'ascolto relativi agli assetti intervallari di questo stile di canto nel quadro di una teoria emica. Per questo motivo la ricerca si è svolta, a questo livello, unicamente in forma sperimentale.

distinzione con altre forme di *mutetu* più brevi e metricamente più semplici, *mutetu longu*). (Bravi, 2008; Zedda, 2008).

3. STRUMENTI E PROCEDURA

La segmentazione dei documenti sonori, l'estrazione della frequenza fondamentale e le operazioni statistiche sui dati sono state effettuate utilizzando il software *Praat* (Boersma, 2001). I documenti sonori analizzati – nel complesso, circa cento – sono tratti da esecuzioni realizzate nel corso di gare poetiche da diverse decine di poeti improvvisatori, con l'eccezione di otto *mutetus* registrati in studio e pubblicati in cassetta verso la fine degli anni '70. Dato che l'esecuzione del *mutetu* comprende anche parti polifoniche, in cui la voce dell'improvvisatore è accompagnata dalle voci di un coro bivocale (detto *basciu e contra*), sono state preliminarmente isolate le parti solistiche. Su tali parti è stata effettuata l'estrazione della frequenza fondamentale, i cui dati sono stati raggruppati in classi con ampiezza di intervallo di 10 cents.[6] Successivamente è stato individuato e isolato con segmentazione manuale il centro tonale (CT).[7] Infine, i dati relativi a ciascuna unità di analisi sono stati rappresentati attraverso diagrammi di frequenza – qui indicati con il termine *tonogramma*, adattato dal francese *tonagramme* (Bel, 1998) e dall'inglese *tonagram* (Van der Meer, 2000)[8] – cui è stata sovrapposta, allo scopo di visualizzare con immediatezza le relazioni intervallari fra le classi, una griglia per semitoni in cui il punto 0 coincide con il centro tonale individuato.

4. RISULTATI

L'esame dei tonogrammi rivela un quadro di stili esecutivi assai variegato. La rilevazione statistica delle frequenze del canto dei poeti improvvisatori mostra comportamenti spesso abbastanza disomogenei, specialmente per quanto riguarda la variazione interindividuale. Lo stile di canto (*tragiu*) dei *cantadoris*, infatti, usualmente mira, se non

[6] Il *cent* è un'unità di misura logaritmica corrispondente a 1/100 di semitono, usata in modo particolare per la misura di piccoli intervalli musicali.

[7] Il termine 'centro tonale' (*tonal center*) è stato introdotto nell'ambito dell'analisi etnomusicologica da Bruno Nettl (Nettl, 1964) con lo scopo di indicare "il riferimento centrale intorno a cui ruota l'elaborazione musicale propria di un *modo* determinato" (Agamennone, 1991: 156). Come osserva Maurizio Agamennone, "[s]i preferisce questa denominazione, piuttosto che quella di *tonica*, poiché quest'ultima espressione, almeno nel lessico italiano, appare strettamente legata alle connotazioni della musica tonale e della *Tonalità*" (Agamennone, 1991: 156). Nel caso presente, l'individuazione del centro tonale è stata realizzata con la seguente procedura: [1] individuazione del livello tonale identificabile come 'centro tonale', effettuata su base percettiva da parte di chi scrive tenendo conto del ruolo strutturalmente centrale del *tonus finalis*; [2] isolamento – realizzato anch'esso su base percettiva da chi scrive – di segmenti di durata consistente in cui il canto si sofferma su tale livello tonale; [3] estrazione di F0 in tali segmenti e calcolo della media dei valori.

[8] Lo stesso termine è usato, in ambito fonetico, per rappresentazioni grafiche di altra natura, che riguardano l'intonazione. Si tratta, in questo caso, di forme di rappresentazione grafica basate sulla distinzione relativa in tre fasce (alta, media, bassa) dell'andamento melodico dell'enunciato (Canepari, 1985; De Dominicis, 1992).

all'originalità, alla riconoscibilità, e rappresenta un elemento che contribuisce in modo significativo alla personalizzazione dello stile poetico nel complesso (Bravi, 2008). Di seguito, mi soffermo su alcuni elementi di particolare rilievo che l'esame dei tonogrammi mette in luce.

4.1 Estensione e gamma

Un primo aspetto che emerge attraverso l'analisi dei tonogrammi è la diversa estensione e il diverso numero di gradi impiegati nelle esecuzioni. Se nella maggioranza dei casi la gamma va dal I al VI grado, vi sono casi in cui l'estensione globale è più ristretta o più ampia, come si può osservare nei due grafici presenti in figura 1.

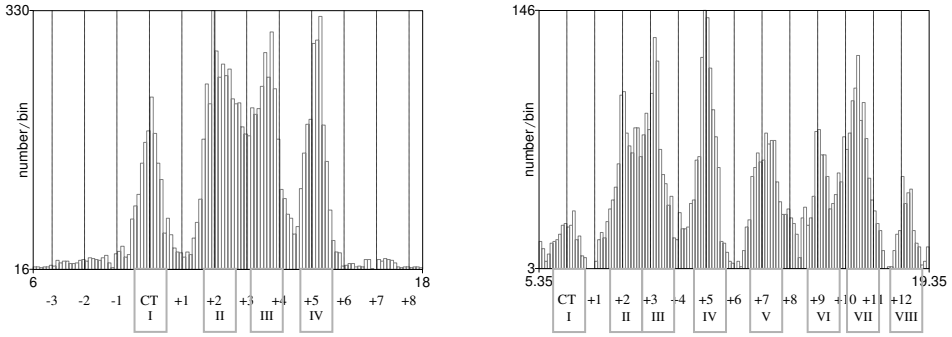


Figura 1: sin., esecuzione del poeta Eliseo Vargiu, con *gamut* di 4 gradi (I-IV); des., esecuzione del poeta Raffaele Cocco, con *gamut* di 8 gradi (I-VIII)[9]

Lo stile di canto di alcuni *cantadoris* si basa su strutture scalari in cui alcuni gradi appaiono, se non del tutto assenti, poco presenti e/o poco definiti.[10] I grafici in figura 2 si riferiscono a esecuzioni in cui alcuni gradi (il II nel primo caso, il II e il III nel secondo caso, indicati e posti in evidenza con riquadri di colore grigio a linee tratteggiate)

[9] La scala delle frequenze (asse *x* del grafico) è logaritmica; l'unità di misura sono i semitoni (0 = 100 Hz), di cui nel grafico sono indicati il valore minimo e massimo. Per rendere più agevole l'individuazione delle relazioni intervallari rispetto al centro tonale individuato, è stata sovrapposta al grafico una griglia verticale che indica la distanza in semitoni dal valore del centro tonale (indicato come CT). Attraverso i riquadri (in grigio) posti sotto l'asse *x* del grafico sono messi in evidenza i punti di concentrazione delle frequenze del canto, interpretati e indicati – da parte di chi scrive – come gradi della scala.

[10] Il riferimento e la definizione numerica dei gradi sono da intendere in casi di questo tipo in modo convenzionale. L'assenza o la presenza limitata o non strutturalmente rilevante di una determinata altezza, sotto il profilo teorico, potrebbe suggerire una diversa 'assegnazione' numerica ai gradi individuati. Per facilitare la lettura dei grafici e per mantenere un legame fra l'ordinamento dei gradi (II, III, IV ecc.) e la definizione dell'ampiezza degli intervalli (seconda, terza, quarta ecc.) rispetto al centro tonale, si è preferito far riferimento alle categorie comuni, ordinariamente usate nell'analisi della struttura delle scale eptatoniche.

sembrano, se non del tutto assenti, piuttosto marginali, almeno per quanto riguarda l'aspetto quantitativo.

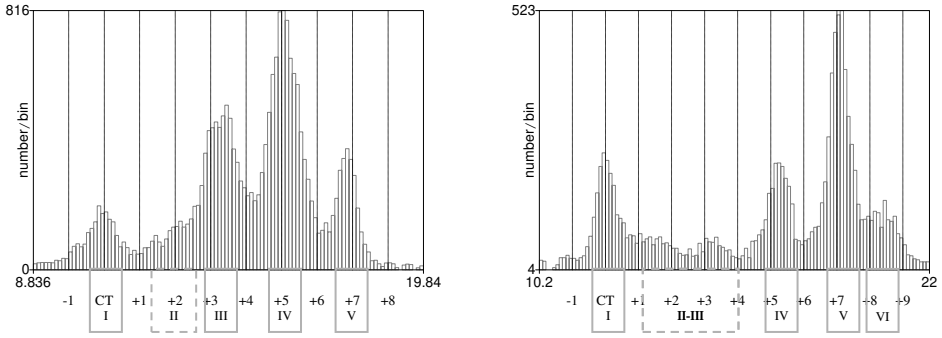


Figura 2: sin., esecuzione del poeta Daniele Filia, con *gamut* incentrato sui gradi I-III-IV-V; des., esecuzione del poeta Federico Lai, con *gamut* incentrato sui gradi I-IV-V-VI

4.2 Evoluzione e variabilità

Gli stili di canto dei *cantadoris* dimostrano, oltre a una notevole varietà inter-individuale, una significativa evoluzione nel corso degli anni. I cambiamenti più marcati si manifestano usualmente nel periodo dell'apprendistato e nei primi anni della carriera ufficiale dei poeti. I tonogrammi nella figura 3 si riferiscono a due esibizioni di un giovane poeta, Pierpaolo Falqui, in un periodo in cui lo stile di canto appare ancora in via di definizione. Si può osservare l'inversione del rapporto fra la frequenza del IV e del V grado – mentre nel 2006 il grado più presente è il IV e il V ha un rilievo minore, nel 2008 si verifica il contrario – e la comparsa, nel 2008, del VI grado, assente nel 2006.

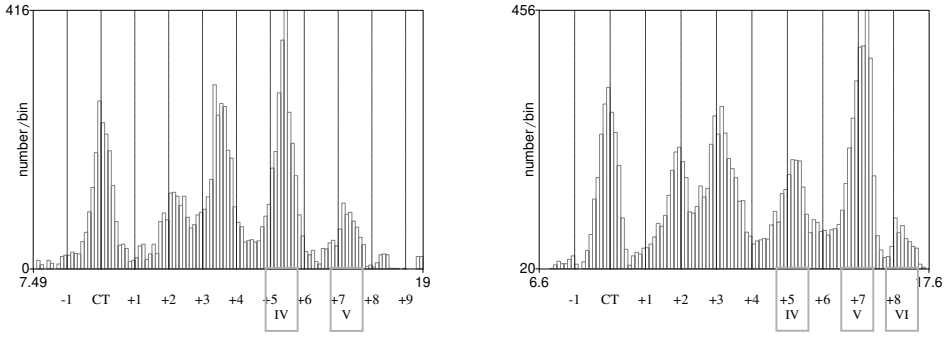


Figura 3: tonogrammi relativi a due esecuzioni del poeta Pierpaolo Falqui, risalenti rispettivamente all'anno 2006 (sin.) e 2008 (des.)

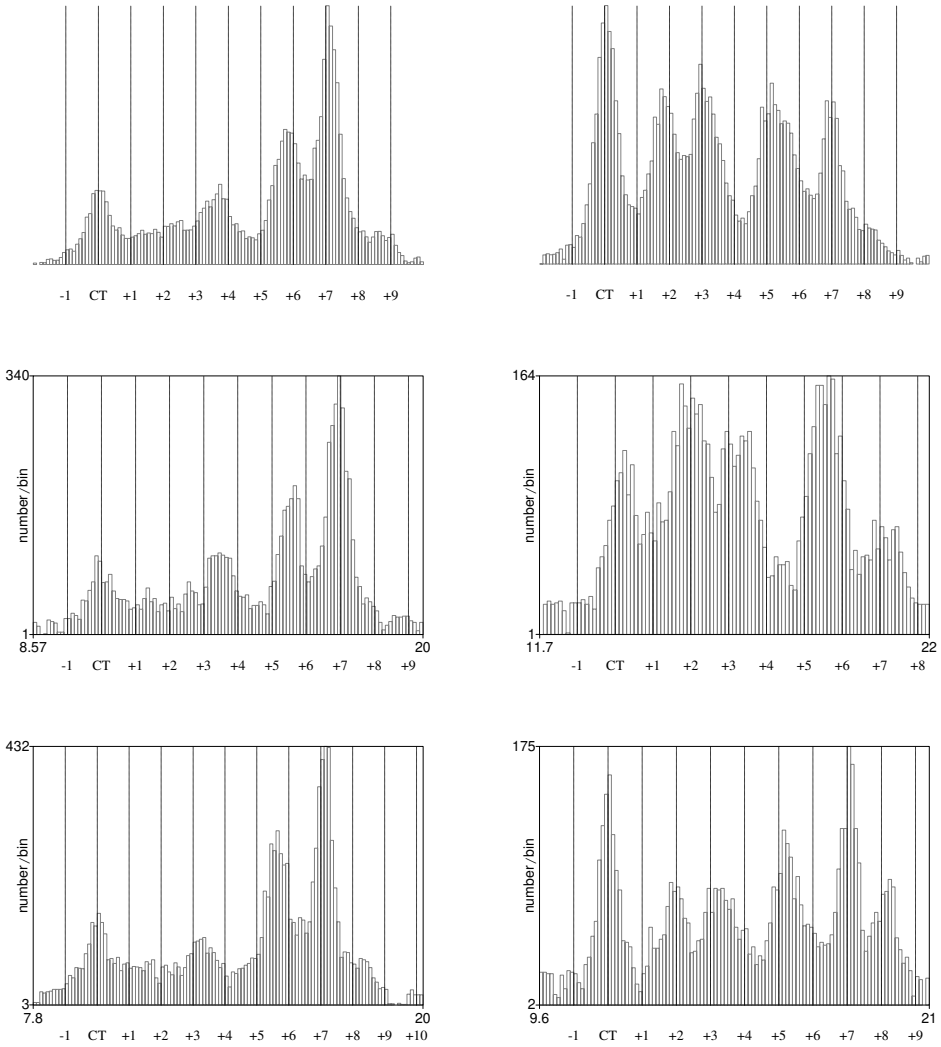


Figura 4: in alto, grafici della distribuzione media delle frequenze in quattro *mutetus* successivi eseguiti rispettivamente dai poeti Ireneo Porceddu (sin.) e Francesco Loddo (des.). Al centro e in basso, grafici relativi al primo e all'ultimo *mutetu* della serie. La distribuzione delle frequenze nei due *mutetus* di Porceddu è assai simile e corrisponde da vicino a quella media, mentre nei due *mutetus* di Loddo è assai diversa e si distacca notevolmente dalla media.

All'interno di una gara poetica, lo stile di canto di ciascun poeta improvvisatore si mantiene in genere relativamente stabile. Tuttavia, sotto questo profilo vi sono differenze fra i *cantadoris*. Alcuni hanno uno stile di canto fondamentalmente omogeneo, altri invece manifestano una variabilità più o meno marcata, che si riflette nella distribuzione delle

frequenze. I tonogrammi in figura 4 si riferiscono all'esecuzione successiva di quattro *mutetus* in una stessa gara poetica da parte di due diversi poeti. I grafici in alto sono ottenuti riportando la frequenza del CT nelle quattro esecuzioni al valore medio. Nella fila centrale e in quella inferiore sono rappresentati i grafici del primo e dell'ultimo *mutetu* preso in

esame. Nel caso rappresentato nel lato sinistro del grafico, che si riferisce al poeta Ireneo Porceddu, i tonogrammi relativi alle due esecuzioni sono assai simili e vicini al grafico che si riferisce ai valori medi; nel caso rappresentato a destra, che si riferisce al poeta Francesco Loddo, il quadro appare invece più variegato e disomogeneo.

4.3 Definizione e dispersione

Un elemento rilevante che emerge dalla comparazione dei grafici è il diverso livello di 'definizione' del grado. In alcuni casi, la "cuspide" risulta ben delineata, con un valore centrale chiaramente individuato; in altri casi invece appare un'area di frequenze in rilievo a volte piuttosto ampia e non ben definita. Oltre a caratterizzare differenze interindividuali, la definizione dei gradi è a volte disomogenea anche nelle esecuzioni di uno stesso poeta improvvisatore. I due grafici in figura 5 rappresentano uno di questi casi. Le due esecuzioni successive del poeta Salvatore Maxia, pur nella sostanziale similarità, evidenziano una diversa dispersione nei gradi II, III e IV.

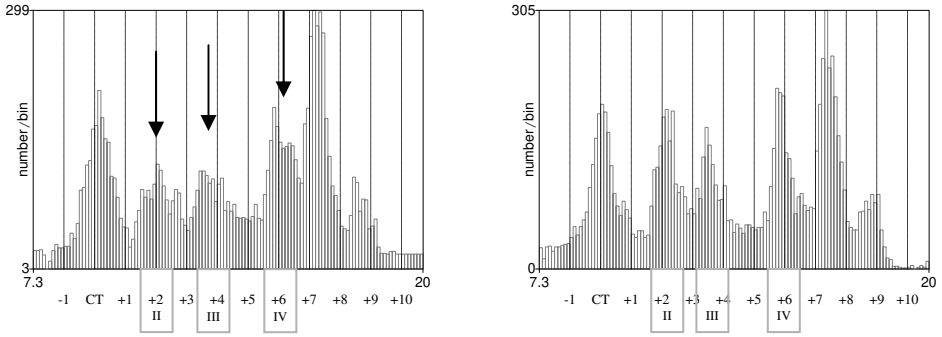


Figura 5: grafici relativi a due esecuzioni consecutive tratte dalla stessa gara poetica (Settimo S. Pietro, 05.08.2003) del poeta Salvatore Maxia. Nel grafico a sin. il II, III e IV grado evidenziano una dispersione maggiore rispetto al grafico a des.

In alcuni casi la dispersione fa sì che i gradi non appaiano delineati e distinti nei tonogrammi. Nei due grafici riportati in figura 6, che si riferiscono a due esecuzioni successive del poeta Giovanni Broi, il II grado (sin.) e il VI grado (des.) si manifestano come 'propaggini' del III e del V grado rispettivamente, piuttosto che come livelli di altezza indipendenti.

La dispersione delle frequenze, secondo gli stili individuali, riguarda spesso alcuni gradi più di altri. In chiave globale, è però possibile osservare questo tipo di fenomeni, sebbene in misura variabile, su tutti i gradi. I grafici in figura 7 evidenziano il fatto che il 'raggruppamento' delle frequenze ascrivibili a gradi congiunti in un blocco fondamentalmente indistinto può riguardare tutte le coppie possibili nella gamma.

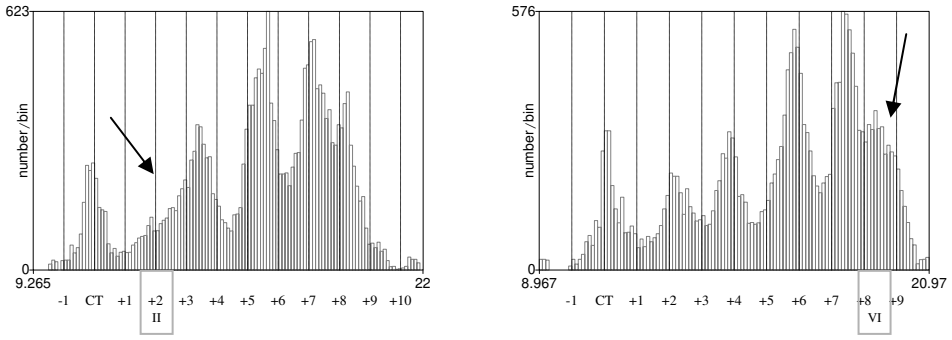


Figura 6: grafici relativi a due esecuzioni consecutive tratte da una registrazione in studio del poeta Giovanni Broi. Il II grado (sin.) e il VI grado (des.) non sono rappresentati nei grafici da cuspidi isolate, ma da aree di frequenza poco definite in adiacenza rispettivamente al III e al V grado

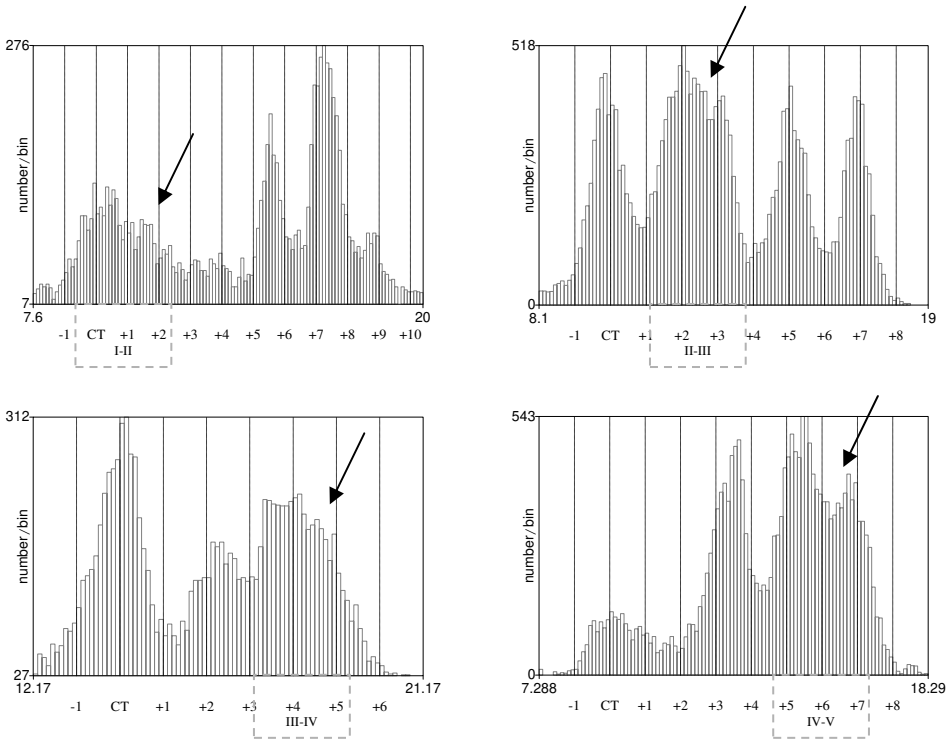


Figura 7: grafici in cui coppie di gradi adiacenti appaiono in tutto o in parte fuse in un ampio blocco di frequenze emergenti. In alto a sin. (Federico Lai), le frequenze del II grado si confondono nel tonogramma con quelle del primo; in alto a des. (Paolo Zedda), II e III grado danno origine ad un tendenziale *plateau* indistinto; in basso a sin. (Francesco Farci), III e IV grado sono rappresentati da un unico gruppo di frequenze emergenti; in basso a des. (Daniele Filia), IV e V grado sono in misura prevalente congiunti.

4.4 Caratteristiche individuali

I tonogrammi rivelano caratteristiche individuali relative all'intonazione nel canto dei poeti e in particolare alle strutture intervallari da essi adottate. Tra le numerose indicazioni che emergono, mi limito ad indicarne alcune.

Tra gli aspetti che caratterizzano lo stile di canto di Salvatore Mascia, un elemento di rilievo riguarda l'intonazione del IV grado. Come si può osservare nei grafici in figura 8, il grado si colloca ad una distanza di circa sei semitoni dal CT, che corrisponde all'intervallo relativamente inusuale di 4ª eccedente.[11]

I grafici in figura 9 mettono in evidenza una caratteristica dello stile di canto di Antonello Orrù, ossia l'intonazione del V grado su un'altezza tendenzialmente inferiore rispetto all'intervallo di 5ª giusta.

I grafici che si riferiscono alle esecuzioni di Salvatore Marras (figura 10) mostrano una tendenza alla equidistanza dei gradi. All'interno dei 'confini' stabiliti dal CT e dal V grado, posto una 5ª giusta sopra il CT, il II, III e IV grado si collocano su altezze tendenzialmente equidistribuite.

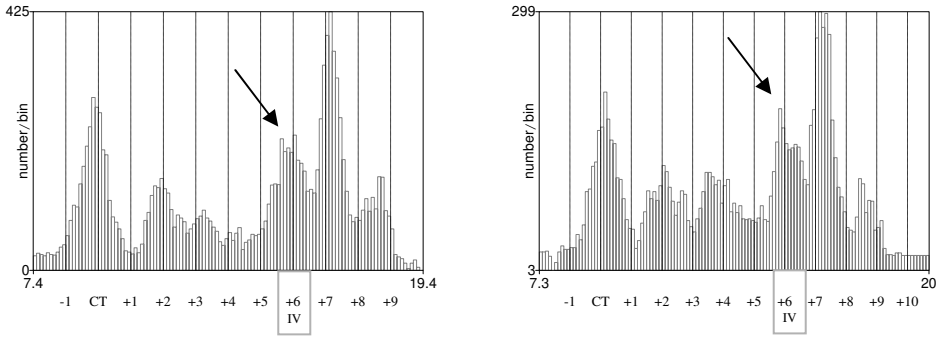


Figura 8: Salvatore Mascia: IV grado vicino all'intervallo di 4ª eccedente

[11] Per i non addetti ai lavori, segnalo che la misura degli intervalli musicali può essere espressa, oltre che in termini quantitativi come distanza in semitoni o *cents*, in termini qualitativi, indicandone "genere" o "nome generico" e "specie" o "nome specifico" (rispettivamente: Giacomoni, 1996: 159; Piston, 1989: 6; a questi titoli rimando anche per le necessarie integrazioni). Do qui una sintetica e parziale indicazione della corrispondenza tra gli intervalli definiti in maniera qualitativa e quantitativa, con riferimento alla scala temperata (vd. *infra*, nota 13): 2ª m. (minore) → 1 st. (semitono); 2ª M. (maggiore) → 2 st.; 3ª m. → 3 st.; 3ª M. → 4 st.; 4ª d. (diminuita) → 4 st.; 4ª g. (giusta) → 5 st.; 4ª e. (eccedente) → 6 st.; 5ª d. → 6 st.; 5ª g. → 7 st.; 5ª e. → 8 st.; 6ª m. → 8 st.; 6ª M. → 9 st.; 7ª m. → 10 st.; 7ª M. → 11 st.; 8ª d. → 11 st.; 8ª g. → 12 st.; 8ª e. → 13 st.

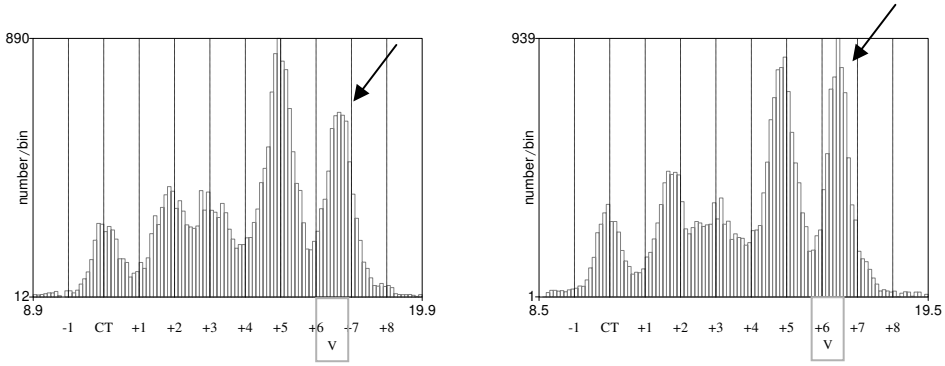


Figura 9: Antonello Orrù: V grado più basso rispetto all'intervallo di 5ª giusta

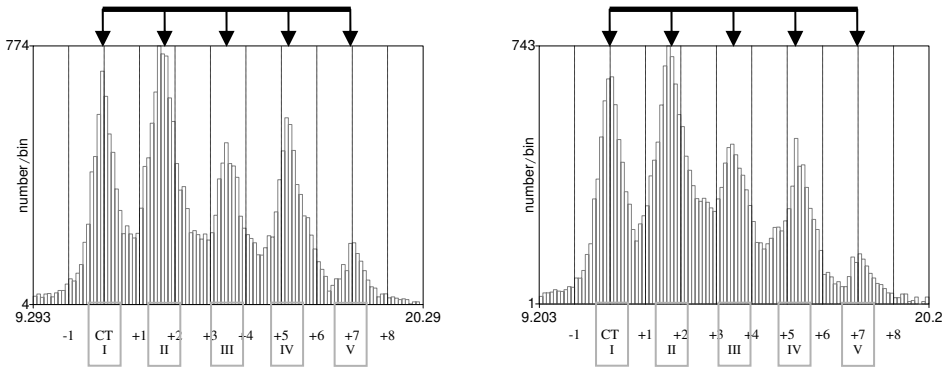


Figura 10: Salvatore Marras: tendenziale equidistanza degli intervalli

5. DISCUSSIONE E CONCLUSIONI

L'interpretazione della distribuzione della frequenza del pitch basata sui tonogrammi pone difficoltà di varia natura.

Alcune di esse riguardano questioni di fondo che è necessario tenere presenti. Fra queste, in primo luogo il fatto che andrebbe presa in considerazione l'ipotesi che l'influenza della dimensione segmentale su F0 osservata nel parlato possa manifestarsi in termini significativi anche nel canto;[12] in secondo luogo, il fatto che i dati dell'analisi acustica devono essere confrontati con i dati percettivi, sia nel caso del parlato (Albano Leoni & Maturi, 1995) sia nel caso del canto e della musica in genere (Aiello & Sloboda, 1994; Sundberg, 1999).

Altri aspetti critici dell'interpretazione dei dati riguardano questioni specifiche. In questo caso, i problemi maggiori sono connessi al fenomeno della dispersione. La presenza di un'ampia dispersione delle frequenze rende difficile in alcuni casi individuare un livello di altezza stabile qualificabile come "grado", definirne i confini e individuarne il valore centrale. Quando la distribuzione delle frequenze assume la forma di una cuspide, si può infatti assumere, almeno in via di ipotesi, di avere di fronte un grado di una scala, la cui altezza corrisponde al picco della cuspide. Nei casi in cui invece la distribuzione delle frequenze appare piuttosto indeterminata, questo non è evidentemente possibile e si pone un problema di interpretazione dei dati. La dispersione delle frequenze, infatti, può essere legata a diversi fattori – cfr (Cohen & Katz, 2005: 43 sgg.) –: può essere connessa a fenomeni di ordine meramente esecutivo (per esempio, a imprecisioni involontarie dell'intonazione da parte del cantore); può dipendere da un uso massiccio di vibrato di notevole ampiezza; può essere collegata a comportamenti intonativi che dipendono da fattori melodici specifici; può rappresentare un elemento intrinseco del sistema musicale, che non prevede una stabilità rigida delle altezze, ecc.

In questo senso, l'esame dei tonogrammi va considerato come un metodo che consente una descrizione *coarse-grained* del sistema di intonazione adottato dai *cantadoris*. La mera rilevazione statistica della distribuzione dei dati non permette di capire ciò che sta dietro a quello che nei diagrammi di frequenza si manifesta come una dispersione. Per analizzare in modo puntuale i comportamenti intonativi dei cantori è invece necessario osservare le variazioni del pitch nei contesti melodici particolari, tenendo in ogni caso presente che la nozione di stabilità delle altezze e il concetto di intervallo spesso non devono essere considerati in modo rigido.

---

[12] A proposito delle relazioni fra la dimensione segmentale e quella intonativa, è significativo osservare che mentre in campo fonetico vi è una mole di studi incentrati sugli effetti della prima sulla seconda nell'ambito del parlato – mi riferisco in particolare al fenomeno dell'*intrinsic pitch*, sul quale una sintetica panoramica è in (Giannini & Pettorino, 1992: 210-213), e a quello delle *obstruent perturbations* (Kingston, 1991) – nel campo dell'analisi del canto l'attenzione si è focalizzata sugli effetti della seconda sulla prima – mi riferisco ad esempio ai fenomeni della *singer's formant* e del *formant matching*, sui quali si trovano indicazioni essenziali in (Sundberg, 2003). È da prendere in considerazione il fatto che il prevalere di una prospettiva o dell'altra, apparentemente legato alla natura dell'oggetto, e cioè al diverso uso e trattamento della voce nel parlato e nel canto, non sia da verificare, almeno in casi specifici, e che non si tratti di un'assunzione di tipo aprioristico.

Preso atto del livello di descrizione che è possibile raggiungere attraverso l'esame dei tonogrammi e dei limiti intrinseci che questo tipo di analisi presenta, si possono comunque trarre alcune conclusioni. Esse potranno servire come "tracce" per indagini più mirate sui comportamenti intonativi dei *cantadoris* campidanesi.

L'analisi dei tonogrammi permette di individuare alcuni elementi di rilievo relativi alle strutture intervallari adottate nello stile del canto *a mutetus*. Emerge con evidenza il fatto che esistono modi assai differenti di cantare, caratterizzati dall'uso di diverse gamme di suoni, dal rilievo diverso dato ai vari gradi, da assetti intervallari diversi, ecc. Ci sono comportamenti intonativi che permettono di intravvedere parentele e derivazioni fra i diversi stili di canto, ma c'è una differenza spesso marcata fra i *cantadoris* e non di rado anche fra esecuzioni diverse da parte di uno stesso *cantadori*. Anche in uno scenario di questo tipo, caratterizzato da forte variabilità, possono peraltro essere osservate alcune linee di tendenza ricorrenti per quanto riguarda gli assetti intervallari. Esse – delineate sinteticamente di seguito e rappresentate nel grafico in figura 11 – riguardano la maggioranza dei poeti e si ritrovano nella gran parte delle esecuzioni.

Il *II grado* si colloca ad una distanza dal CT che è in genere inferiore all'intervallo di 2ª maggiore (in alcuni casi appare inferiore al semitono). Il *III grado* manifesta un'ampia variabilità – anche all'interno di singole esecuzioni – e si colloca spesso ad una distanza intermedia fra gli intervalli di 3ª minore e 3ª maggiore dal CT. La collocazione prevalente è più vicina o coincidente con l'intervallo minore, ma non sono affatto rari i casi in cui l'altezza è invece prossima all'intervallo di 3ª maggiore. Il *IV grado*, talvolta posto una 4ª giusta sopra il CT, si colloca invece nella maggioranza dei casi ad una distanza tendenzialmente superiore, fino ad arrivare ad una distanza di circa tre toni dal CT. Il *V grado*, pur con lievi oscillazioni verso l'acuto o verso il grave, si trova mediamente una 5ª giusta sopra il CT. Il *VI grado*, infine, che ha una presenza limitata e un rilievo quasi sempre secondario, manifesta caratteristiche simili a quelle del III grado: in molti casi si trova ad una distanza vicina all'intervallo di 6ª minore dal CT, ma non sono isolati i casi in cui l'intervallo è invece più vicino ad una 6ª maggiore o occupa una posizione intermedia.

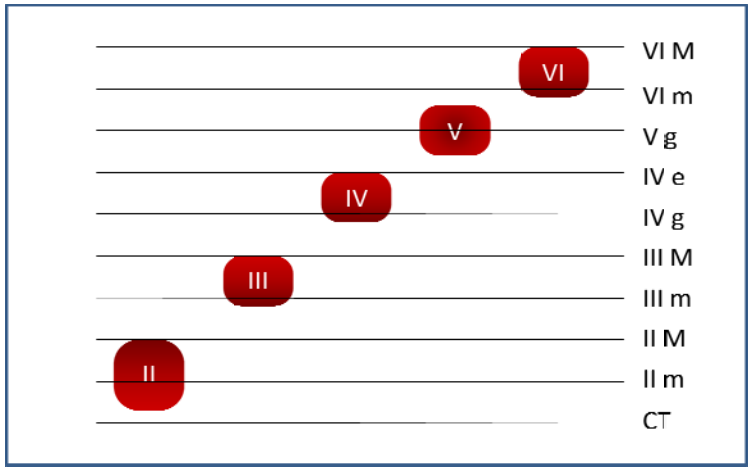


Figura 11: Linee di tendenza nella definizione degli assetti intervallari nel canto *a mutetus*

Una questione che si pone è se il sistema intervallare, pur nella sua limitata regolarità e omogeneità, abbia manifestato in tempi relativamente recenti un mutamento in direzione della scala temperata, che costituisce il sistema su cui si fonda gran parte della musica occidentale moderna.[13] Sebbene l'analisi della distribuzione delle frequenze offra indicazioni non univoche e in alcuni casi del tutto contraddittorie, in almeno alcuni fra i *cantadoris* delle generazioni recenti sembra manifestarsi una tendenza ad un avvicinamento alla strutturazione prevista dal sistema temperato.[14] In particolare, in diversi casi le altezze tendono ad assumere una configurazione vicina a quella definita dalla scala minore, come si può osservare nei grafici seguenti (figura 12), che riguardano due *mutetus* rispettivamente di Eliseo Vargiu (classe 1959) e Pierpaolo Falqui (classe 1981).

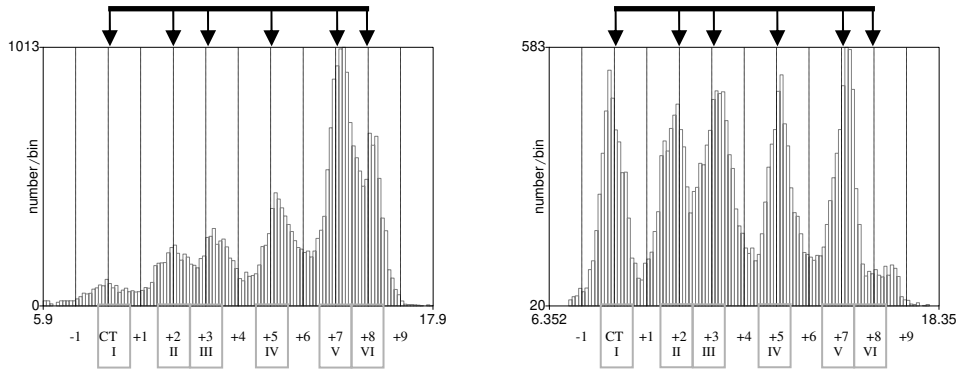


Figura 12: gli assetti intervallari nelle esecuzioni di alcuni *cantadoris* sono vicini a quelli del modo minore. A sin., Eliseo Vargiu; a des. Pierpaolo Falqui.

È infine opportuno richiamare un punto importante per i possibili sviluppi futuri di questa indagine. Gli assetti intervallari del canto – così come le strutture musicali in genere – rivestono un valore sociale in relazione al contesto in cui essi prendono vita. Gli elementi che costituiscono un sistema musicale, infatti, devono essere considerati come "fatti

[13] Per i non addetti ai lavori, segnalo che con l'indicazione "scala (cromatica) temperata" (o temperamento equabile) si indica la scala in cui l'intervallo di ottava è suddiviso in dodici intervalli uguali, in cui il rapporto di frequenza del semitono (temperato) è $\sqrt[12]{2}$. Per quanto riguarda la tradizione musicale 'classica' occidentale, la scala temperata rappresenta un'evoluzione radicale rispetto ad altri sistemi scalari definiti in precedenza, come la scala pitagorica e la scala zarliniana o naturale (Righini, 1989).

[14] Secondo Paolo Zedda nel periodo a cavallo tra gli anni Cinquanta e Sessanta "radio e televisione impongono il modello occidentale della scala temperata che praticamente monopolizza il sistema mediatico. Da quegli anni in poi il cervello dei sardi si forma con questo sistema di impostazione melodica continuamente ribadito, per ore ogni giorno, da radio, dischi e televisione, e percepito come riferimento 'corretto', con la conseguenza che gli altri sistemi melodici risultano "stonati" o grezzi e imprecisi". La conseguenza di questo 'bombardamento' musical-mediatico è l'"abbandono della impostazione scalare autoctona in favore di quella temperata" (in Cirese *et al*., 2006: 26).

sociali" che assumono significato all'interno di paradigmi di interpretazione culturalmente definiti e variabili. Come osservava ormai un quarto di secolo fa John Blacking:

> […] not only do individuals and groups give different verbal meanings to music; they also conceive its structures in ways that do not permit us to regard musical parameters as objective acoustical facts. In music, thirds, fourths, fifths, and even octaves, are social facts, whose syntactical behaviour can differ as much as that of *si*, *see*, and *sea*, *beau*, *bow*, and *bo*, or *buy*, *bye*, *by* and *bai*. (Blacking, 1984: 364)

Tra gli elementi che inducono a valutare la presente ricerca come un passo iniziale, vi è dunque anche la necessità di considerare il fatto che gli assetti intervallari utilizzati dai *cantadoris*, osservati in questo lavoro esclusivamente nei loro caratteri acustici, sono elementi di base di un sistema musicale la cui valutazione e interpretazione almeno in parte trascende la sfera dei dati oggettivi e misurabili, ed ha invece a che fare con il modo con cui il cantare dei *cantadoris* crea significati, genera riflessioni e alimenta discorsi.

6. BIBLIOGRAFIA

Agamennone, M. (1991), Modalità / Tonalità, in *Grammatica della musica etnica* (M. Agamennone, S., Facci, F., Giannattasio & G., Giuriati, editors), Roma: Bulzoni, 145-200.

Aiello, R. & Sloboda, J. A., editors (1994), *Musical Perceptions*, New York: Oxford University Press.

Albano Leoni, F. & Maturi, P. (1995), *Manuale di fonetica*, Roma: La Nuova Italia Scientifica.

Baily, J. (2005), La teoria musicale nelle tradizioni orali, in *Enciclopedia della musica. Vol. V: L'unità della musica*, Torino: Einaudi, 537-554.

Bel, B. (1998), Raga: approches conceptuelles et expérimentales, in *Actes du Colloque "Structures Musicales et Assistance Informatique",* CRSM-MIM, Aix-en-Provence / Marseille, 87-108.

Blacking, J. (1984), What languages do musical grammars describe?, in *Musical Grammars and Computer Analysis* (M. Baroni and L. Callegari, editors), Firenze: Olschki, 363-370.

Boersma, P. (2001), Praat, a system for doing phonetics by computer, in *Glot International* 5(9/10), 341-345.

Bravi, P. (2008), A sa moda campidanesa*. Pratiche, poetiche e voci degli improvvisatori nella Sardegna meridionale*, Tesi di Dottorato, Università di Siena, Italy.

Canepari, L. (1985), *L'intonazione. Linguistica e paralinguistica*, Napoli: Liguori.

Cardona, G. R. (1985), *La foresta di piume. Manuale di etnoscienza*, Roma-Bari: Laterza.

Cirese, A. M., Murru, A. & Zedda, P. (2006), Unu de Danimarca benit a carculai. *Il mondo poetico di Ortacesus nelle registrazioni e negli studi di Andreas Fridolin Weis Bentzon tra il 1957 e il 1962*, Cagliari: Iscandula.

Cohen, D. & Katz, R. (2005), *Palestinian Arab Music. A Maqam Tradition in Practice*, Chicago: University of Chicago Press.

De Dominicis, A. (1992), *Intonazione e contesto. Uno studio su alcuni aspetti del discorso in contesto e delle sue manifestazioni intonative*, Alessandria: Edizioni dell'Orso.

Ellis, A. J. (1885), On the Musical Scales of Various Nations, *Journal of the Society of Arts*, 33, 485-527.

Ferrari, F. (ed) (1999), *Scrivere la musica*: *per una didattica* delle notazioni, Torino: EDT.

Giacomoni, G. (1996), *Elementi di teoria musicale*, Parma: Azzali.

Giannattasio, F. (2005), Dal parlato al cantato, in *Enciclopedia della musica. Vol. V: L'unità della musica*, Torino: Einaudi, 1003-1036.

Giannini, A. & Pettorino, M. (1992), *La fonetica sperimentale*, Napoli: Edizioni Scientifiche Italiane,

Giuriati, G. (1991), Trascrizione, in *Grammatica della musica etnica* (M. Agamennone, S. Facci, F. Giannattasio & G. Giuriati, editors), Roma: Bulzoni, 243-290.

Hirst, D. J., Di Cristo, A., editors (1998), *Intonation Systems. A Survey of Twenty Languages*, Cambridge: Cambridge University Press.

Hornböstel, E. M. (1913), Melodie und Skala, *Jahrbuch der Musikbibliothek Peters*, 19, 11-23.

Kingston, J. (1991), Integrating articulations in the perception of vowel height, *Phonetica*, 47, 149-179.

List, G. (1963), The boundaries of speech and song, *Ethnomusicology*, VII, no. 1, 1-16.

Nattiez, J. J. (1981), Scala, in *Enciclopedia*, vol. XII, Torino: Einaudi, 454-470.

Nettl, B. (1964), *Theory and Method in Ethnomusicology*, Glencoe: Free Press.

Patel, A. (2008), Talk of the tone, *Nature*, 453, 726-727.

Pierrehumbert, J. & Beckman, M.E. (1988), *Japanese Tone Structure*, Cambridge, MA: MIT Press.

Piston, W. (1989), *Armonia,* Torino: EDT.

Powers, H. S. (1980), Mode, in *The New Grove Dictionary of Music and Musicians*, vol. XII, 376-450.

Powers, H. S. (1992), Modality as a European cultural construct, in *Secondo convegno europeo di analisi musicale* (R. Dalmonte & M. Baroni, editors), Trento: Università degli Studi di Trento, 207-220.

Righini, P. (1989), *Il temperamento*, Padova: Zanibon.

Sloboda, J. A. (1988), *La mente musicale. Psicologia cognitivista della musica*, Bologna: Il Mulino.

Sorce Keller, M. (1990), Alcune considerazioni per uno studio analitico della melodia nelle trascrizioni popolari e in quelle extraeuropee, appendice a, *Analisi della struttura melodica* (M. De Natale, editor), Milano: Guerini e ass., 209-229.

Sundberg, J. (1999), The Perception of Singing, in *The Psychology of Music* (D. Deutsch, editors), 2nd edition, San Diego: Academic Press, 171-214.

Sundberg, J. (2003), Research on the singing voice in retrospect, *Speech, Music and Hearing Quarterly Progress and Status Report (TMH-QPSR)*, Vol. 45, 11-22.

Tjernlund, P., Sundberg, J. & Fransson, F. (1972), Grundfrequenzmessungen an schwedischen Kernspaltflöten, *Studia Instrumentorum Musicae Popularis*, 2, 77-96.

Van der Meer, W. (2000), Theory and Practice of Intonation in Hindusthani Music, in *The Ratio Book* (C. Barlow, editor), Köln: Feedback Papers, 50-71.

Zedda, P. (2008), *L'arte de is mutetus. Il canto e l'improvvisazione nei poeti sardi del Campidano*, Iesa: Edizioni Gorée.

PERCEZIONE
E APPRENDIMENTO

FUNCTIONS OF THE LEFT AND RIGHT POSTERIOR TEMPORAL LOBE DURING SEGMENTAL AND SUPRASEGMENTAL SPEECH PERCEPTION

Cyrill Ott, Martin Meyer
Institute of Neuropsychology, University of Zurich
c.ott@psychologie.uzh.ch, m.meyer@psychologie.uzh.ch

1. ABSTRACT

This manuscript reviews evidence from neuroimaging studies on elementary processes of speech perception and their implications for our understanding of the brain-speech relationship. Essentially, differential preferences of the left and right auditory-related cortex for rapidly and slowly changing acoustic cues that constitute (sub)segmental and supra-segmental parameters, e.g. phonemes, prosody, and rhythm. The adopted parameter-based research approach takes the early stages of speech perception as being of fundamental relevance for simple as well as complex language functions. The current state of knowledge necessitates an extensive revision of the classical neurologically oriented model of language processing that was aimed at identifying the neural correlates of linguistic components (e.g. phonology, syntax and semantics) more than at substantiating the importance of (supra)segmental information during speech perception.

2. BACKGROUND

More than a century ago the fundamental discoveries of Paul Broca (1863) and Carl Wernicke (1874) demonstrated that lesions in anterior and posterior perisylvian parts of the left hemisphere cause major impairments of speech functions which were not encountered after focal damage to homotopic areas of the right hemisphere. This has led to the still widely held belief that in most individuals language is an exclusive capacity of the left hemisphere (LH) whereas the right hemisphere (RH) is only marginally involved in language processing. In particular, the classical neurological model associated all essential expressive and perceptive aspects of language with two anatomically ill-defined regions – 'Broca's' and 'Wernicke's area' – in the left perisylvian cortex that are even at present frequently mislabeled 'speech centers' ('Sprachzentren') (Birbaumer & Schmidt, 2002) or considered "specialized (...) delimited regions" that "seemed to be organized explicitly for the processing of verbal information" (Geschwind, 1979). In the sixties of the 20th century Geschwind modified this scheme by proposing an additional area in the inferior parietal lobe that primarily mediates semantic functions (Ben Shalom & Poeppel, 2008). The view that there might exist one or more principal 'speech areas' in the human brain received lively support when Galaburda and colleagues published their seminal paper on a cytoarchi-tectonically distinct temporoparietal area Tpt in the mediocaudal supratemporal plane (Galaburda, Sanides & Geschwind, 1978). Akin to the planum temporale, the area Tpt features a marked leftward asymmetry which made the authors reason that "it may represent, at least in part, the anatomic substrate for language lateralization" (812).

However, with the advent of neuroimaging techniques almost three decades later modifications became inevitable. Despite of all methodological limitations and technical constraints (Brett, Johnsrude & Owen, 2002) and notwithstanding the concerns raised by

some researchers admonishing the lack of a solid theoretical foundation and appropriate concepts of imaging research (Poeppel & Embick, 2005; Sidtis, 2007; Van Lancker Sidtis, 2006) the abundance of recent imaging work on the brain-language relationship has provided at least two basic insights. First, the notion of distinct 'speech centers' residing in the human brain has been falsified. Rather numerous intertwined peri- and extrasylvian regions partake in a variety of linguistic and paralinguistic functions. In particular, 'Broca's area', the ventral most dip of the left inferior frontal gyrus (IFG), that has for long been considered the site of grammatical functions (Dapretto & Bookheimer, 1999; Embick, Marantz, Miyashita, O'Neil & Sakai, 2000; Grodzinsky, 2000) is now viewed as a universal processing device that plays an eminent role in a number of speech and non-speech tasks, such as detecting structural properties, interpreting hierarchical organization, observing and imitating orofacial actions, and recognizing dependencies between repeated or related elements independent of particular cognitive domains (Binkofski & Buccino, 2004; Fadiga, Craighero & Roy, 2006; Fink, Manjaly, Stephan *et al.*, 2006; Friederici, 2006; Hoen, Pachot-Clouard, Segebarth & Dominey, 2006; Koechlin & Jubault, 2006; Marcus, Vouloumanos & Sag, 2003; Meyer & Jancke, 2006; Tettamanti & Weniger, 2006). Furthermore, there is also considerable inconsistency with respect to the precise anatomical definition of 'Broca's area' (Amunts & Zilles, 2006; Petrides, 2006) and the attempt to bridge "the gap between the anatomically based term 'Broca's area' and the increasing number of subfunctions attributed to this area is arbitrary" (Lindenberg, Fangerau & Seitz, 2007: 22). Undoubtedly, the lateral convexity of the left IFG is involved in the processing of structural features of human language (Friederici, 2004, 2006; Friederici, Bahlmann, Heim, Schubotz & Anwander, 2006; Grodzinsky, 2006; Grodzinsky & Friederici, 2006), but it has become rather evident that it also subserves hierarchical and sequential processes in general and should be considered a "higher-level modality-independent control center for executive processes" (Lindenberg *et al.*, 2007: 27).

A considerable lack of consensus also exists with respect to precise size and location of "Wernicke's area" which has been taken to be a speech-selective region for a long time (Bogen & Bogen, 1976; Wise, Scott, Blank *et al.*, 2001). Depending on the varying definitions, 'Wernicke's area' is loosely assigned to the posterior part of the left perisylvian cortex, partly covering the planum temporale (PT), the lateral convexity of the superior temporal gyrus (STG), the most posterior and medial part of the supratemporal plane (STP) at the junction with the inferior parietal lobe (IPL), and the upper bank of the posterior superior temporal sulcus (STS). A large number of neuropsychological and neuroimaging studies have associated quite different speech-related but also non-speech auditory functions with one or several of these anatomically defined areas (Griffiths & Warren, 2002; Hickok & Poeppel, 2004; Wise *et al.*, 2001). The arbitrary use of the term 'Wernicke's area' in association with a large variety of language tasks has led to notable confusion. It therefore appears reasonable to abandon the notion of 'Wernicke's area' and much will be gained by investigating the sensitivity of the aforementioned anatomically circumscribed perisylvian regions separately with respect to speech perception. In the context of the present review I prefer to use the general and rather loose term 'posterior auditory-related cortex' (pARC) as suggested by Hackett & Kaas (2004) when referring to the entirety of aforementioned regions. The pARC denotes an array of bilateral adjacent territories along the Sylvian fissure and beyond, namely Heschl's gyrus (HG), the PT, the planum parietale, the parietal operculum, and the STS that have been more or less tightly

associated with elemental aspects of auditory and speech perception. Furthermore, since the terms 'Broca's area' and 'Wernicke's area' only refer to anatomical sites in the left hemisphere I will not use them in the context of this article as there is compelling evidence that patches of right perisylvian cortex also mediate auditory information processing and hence should be considered an integral part of the neural circuits that subserve elementary as well as complex aspects of speech (Crinion & Price, 2005; Poeppel, Guillemin, Thompson *et al.*, 2004; Uppenkamp, Johnsrude, Norris, Marslen-Wilson & Patterson, 2006).

Hence, the present article puts a particular emphasis on the perceptive and cognitive contributions of the right brain-half considered to be the nondominant hemisphere with respect to language processing in the vast majority of individuals (due to Price (2000) 95% of right-handed and approximately two-thirds to three-quarters of left-handed individuals).[1]

The classical neurological model of the speech-brain relationship holds the lopsided view of speech being lateralized in the dominant hemisphere (Price, 2000). However, the relevance of the right 'step-hemisphere' must not be underestimated (Ben Shalom & Poeppel, 2008; Jung-Beeman, 2005; Poeppel & Embick, 2005; Poeppel & Hickok, 2004; Stowe, Haverkort & Zwarts, 2005). Meanwhile an increasing number agree in that some long-held beliefs must be discarded; the classical neurological model needs substantial revision as it [a] cannot account for the entire range of impairments found in language processing, [b] is restricted to the notion of linguistic components (semantics, syntax, phonology, prosody), ignoring the interplay between them and is therefore underspecified with respect to language as a means of communicative interactions, [c] has led to false assumptions on 'speech-selective' regions in the left hemisphere, and is [d] anatomically underspecified (Ben Shalom & Poeppel, 2008). Alternatively, Ben Shalom and Poeppel propose a model of the functional neuroanatomy of language in which the three main linguistic components (syntax, semantics, phonology) are tied up with the relevant type of processing found in a given cortical area (memorizing in temporal cortex, analyzing in parietal cortex, and synthesizing in frontal cortex). Here, we do not seek to relate specific linguistic components with a particular brain site, but our emphasis is on the elementary sublexical stages of speech perception and we will expound how these computational steps may be represented in the brain.

The present review is focused on more recent models of speech perception which argue more strongly for a parameter-based (rather than component-based) approach. This is an important issue because the majority of neuroimaging studies that have been published during the last decade have attempted to identify the brain mechanisms underlying higher linguistic entities, namely syntax, semantics and phonology. However this abundance of studies has failed to draw a coherent picture of the neural substrates subserving syntax and semantics. According to recent reviews and meta-analyses the entire left (and right) perisylvian cortex is involved in syntactic, semantic and phonological perception and comprehension (Demonet, Thierry & Cardebat, 2005; Grodzinsky & Friederici, 2006; Kaan & Swaab, 2002; Vigneau, Beaucousin, Herve *et al.*, 2006) with no particular site evidently mediating a particular linguistic component. This finding comes as no surprise as

[1] However, it should be mentioned that often the mere activation of the left inferior frontal cortex is taken to index language dominance while the recruitment of the posterior auditory cortex is ignored (Pujol, Deus, Losilla & Capdevila, 1999).

the question whether linguistic components may reside in the human brain, is evidently the wrong question. It rather seems more appropriate to revert to the elemental perceptual categories and to depart from a task-centered approach that has no neurobiological equivalence and therefore lacks conceptual plausibility. In other words, it appears to be more promising to examine how physical (acoustic) signals that encode linguistically relevant (segmental und suprasegmental) information are recognized by the auditory cortex and how they are transformed into representations that are used for linguistic computations in the brain. While the latter is still not well understood the investigation of the former has thus far provided revealing insights which also shed a new light on the cerebral lateralization of brain functions.

Before sketching the parameter-based approach in more detail (cf. 2.1) and introducing recent evidence that corroborates this approach (cf. 2.2) we want to point out two issues that will not be addressed in this review. First, this review is mainly concerned with the brain mechanisms of language processing in the auditory domain. There is a broad consensus that reading and writing are organized differently, both conceptually and with respect to the underlying neural substrates. Readers who are interested in the neural underpinnings of the disorders in written language processing may find an account in the recent work by Price and colleagues (Price, 2000; Price, Gorno-Tempini, Graham *et al.*, 2003; Price & Mechelli, 2005).

Secondly, as mentioned above both the anterior and the posterior part of the Sylvian fissure have been associated with speech functions. The present review places particular emphasis on the posterior part that harbors the posterior auditory-related cortex whereas the functions of inferior frontal regions, namely "Broca's area" and the adjoining frontal operculum in the context of speech will be addressed only marginally. A number of recent publications have elaborated on this issue comprehensively, amongst others a special issue of *Brain and Language* (Volume 89 (2), 2004), a special issue of *Cortex* (Volume 42 (4), 2006), and a text book edited by Yosef Grodzinsky and Katrin Amunts ('Broca's region', Oxford University Press, 2006).

2.1 Segmental and suprasegmental information processing in the auditory domain

Spoken language is an acoustic signal which unfolds in time. Thus, speech comprises acoustic information that may be described at the frequency and at the time scale level. For example, a spoken sentence is characterized by acoustic cues that change rapidly within brief segments. All the frequency variations that mark voicing (e.g. formant transitions, voice onset times, consonantal bursts) have a temporal grain at a rate of milliseconds and occur slowly in a suprasegmental mode (e.g. intonation contour) that unfolds at a rate of hundreds or even thousands of milliseconds (Phillips & Farmer, 1990). Thus, while the former variations constitute small computational units that, for example, encode either voiced or voiceless consonants, the latter span over several segments and thus constitute prosodic phenomena, (e.g. sentence mode, word accent) that are thus similar to musical melodies.

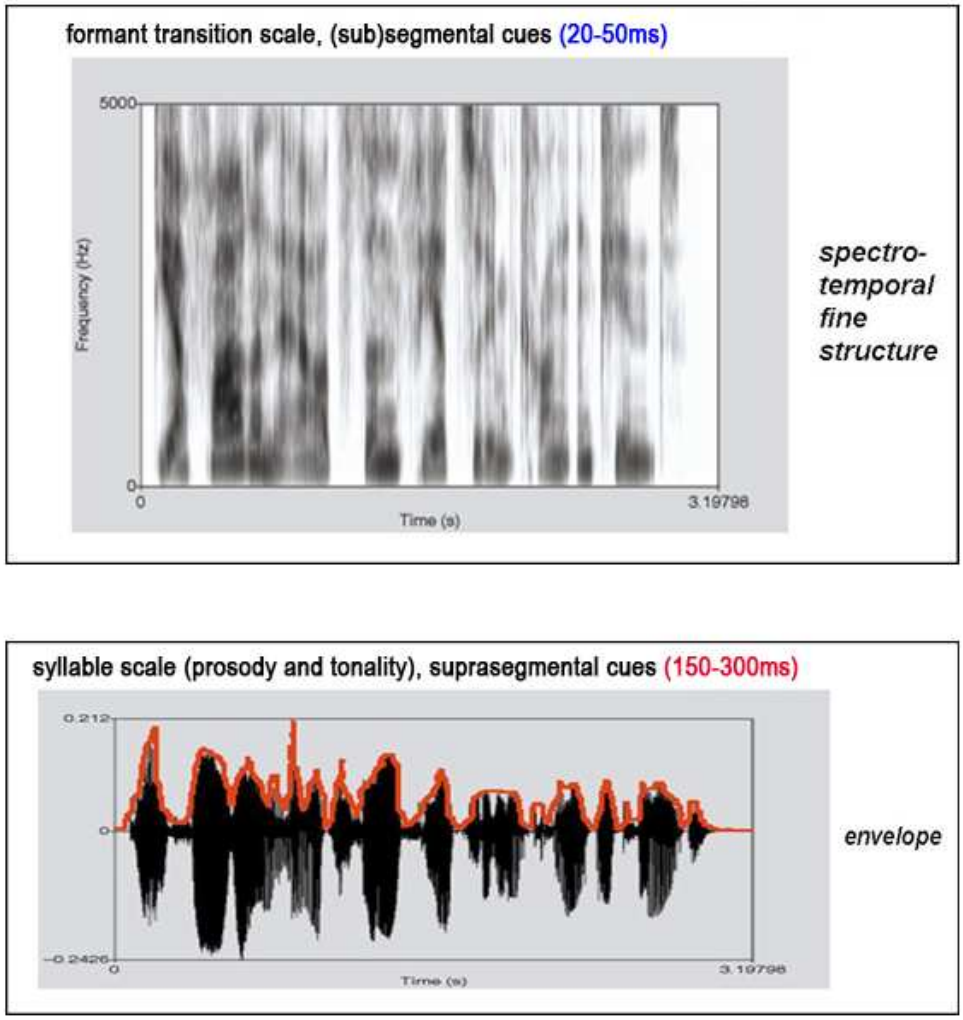


Figure 1: Spectro-temporal information available in speech sounds[2]

Complementary, spectral cues (frequency) encode information pertaining to pitch height, timbre and tone that also characterizes a number of important acoustic and perceptive qualities of speech signals. Spectro-temporal acoustic information (and their

[2] The upper image shows the spectrogram of the German sentence *Der erfahrene Arzt fährt zu dem kranken Kind* ("The experienced doctor visits the sick child."). Frequency changes in both the formant and harmonic structure are typical for vocalization and speech. The lower image illustrates a waveform of the same sentence. Distances between distinct sound bursts and amplitude envelope of bursts highlighted by the red line indicate the rhythmical structure and intonation contour of a spoken utterance.

modulations) should hence be considered the foundation of the acoustic processing of speech, but also of non-speech auditory signals.

During the last decade fruitful research has been carried out to reconcile the concept of computational time windows of different duration with a 'division of labor' between the left and the right auditory cortex. Briefly, based on these concepts it is proposed that left hemispheric specialization for speech is a result of asymmetries in basic auditory processing. According to this view neural ensembles in the left posterior auditory-related cortex are preferentially driven by rapidly changing acoustic cues, namely formant transitions. A left hemisphere preference for rapidly modulating auditory information in the context of speech regardless of linguistic content has been observed more than forty years ago for the first time (Efron, 1963; Poeck & Pietron, 1981; Tallal, Miller & Fitch, 1993). Obviously, rapidly changing cues are an inherent feature of spoken language but it is not the linguistic segment per se that prompts left pARC involvement. This reasoning is supported by the observation that deficient processing of elemental auditory information, rather than a deficit in linguistic abilities, could account for phonetic processing disorders observed in neurological patients and language-learning impaired children (Schwartz & Tallal, 1980; Tallal, Miller, Bedi *et al.*, 1996; von Steinbuchel, 1998). In other words, speech perception relies on intact mechanisms of time-resolution at a time scale level of milliseconds. In accordance, seminal behavioral work has demonstrated that intact recognition of speech is still possible when spectral cues are almost completely removed from the speech signal while temporal and amplitude cues are preserved in each spectral band (Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995). Together, this evidence points to the existence of a universal cortical system mediating auditory temporal resolution in speech and non-speech sounds.

However, one behavioral study demonstrated that the intelligibility of auditory sentences is preserved when local segments of a spoken utterance are presented in reversed order (Saberi & Perrott, 1999). Due to the authors rapidly changing cues cannot be considered the sole key to intelligibility. Rather it seems that "ultralow-frequency modulation envelopes in the order of 3 to 8 Hz are the critical cues to intelligibility" (760). Interestingly the time range around 4 Hz has also been reported as essentially important for speech perception (Poeppel, 2001, 2003). I will get back to this issue later in the article.

According to of Zatorre and co-workers (Zatorre & Belin, 2001; Zatorre, Belin & Penhune, 2002) symmetries in auditory processing may be considered the developmental outcome of optimizing the processing of acoustic cues, with left auditory cortical areas being highly proficient at temporal resolution. Complementarily, right auditory cortical areas are more amenable to spectral resolution. Along this line of argument, a leftward asymmetry in response to rapid frequency transitions (~ 40 ms) was reported in a PET-study when participants heard nonverbal sounds (Belin, Zilbovicius, Crozier *et al.*, 1998). Slow frequency transitions (~ 200 ms) were found to be processed in both the left and the right auditory cortex. In two follow-up studies volunteers were presented with pure tones that varied in the temporal and spectral domains (Jamison, Watkins, Bishop & Matthews, 2006; Zatorre & Belin, 2001). In essence, responses to the temporal features were weighted towards the left pARC, while responses to the spectral features were weighted towards the right pARC. The authors concluded that the results support the notion of a complementary hemispheric specialization of auditory-related cortex for temporal (left) and spectral information (right) in both speech and non-speech sounds. A recent MEG study of

Okamoto *et al.* (2009) used similar stimuli (pure tones and tone pulse trains), which varied over time in either the spectral or temporal dimension in order to obtain processing preferences of the two hemispheres at a higher time-resolution level. The results obtained in this study confirmed that neural responses elicited by spectral versus temporal changes differed between hemispheres relatively early (at about 100 ms), with temporal changes evoking significantly higher N1m responses over the left, and spectral changes eliciting higher N1m responses over the right auditory related cortex. Given the assumption that the right hemisphere exhibits more clearly spectrally organized tonotopic maps than the left (as shown by Liégois-Chauvel *et al.*, 2001), spectral information would preferentially be represented tonotopically within the right hemisphere. On the other hand, evidence from intracerebral evoked potentials suggests that processing of fine-grained durational properties of auditory input is localized to the left Heschl's Gyrus and planum temporale (Liégois-Chauvel, deGraaf, Laguitton & Chauvel, 1999), implying that temporal information would be preferentially encoded by the left hemisphere into a temporal pattern of corresponding neural activity, which perfectly fits the results of Okamoto *et al.* as well as those of several recent neuroimaging studies (e.g. Jamison *et al.*, 2006; Belin *et al.*, 1998; Zaehle *et al.*, 2004). However, it should be mentioned that there exists some evidence indicating this division between hemispheres might not be so clear-cut as these findings suggest (e.g. Schonwiesner, Rubsamen & von Cramon, 2005). Furthermore, Okamoto and coworkers concluded in line with the aforementioned non speech-specific asymmetry, that "the hemispheric laterality of neural responses does not depend on a specific sound type but rather on spectral and temporal variances, which are differentially encoded into neural activity in the auditory related cortex" (5). Thus, hemispheric lateralization of auditory processing is not limited to sounds that carry meaning (as speech or music) but "at least partly originates from early basic neural processing levels dealing with the spectral and temporal features of auditory inputs" (5). However, comparisons of the N1m-latencies further revealed significantly delayed N1m-responses elicited by temporal changes compared to spectral changes, supporting the hypothesis of different neural mechanisms underlying the processing of temporal and spectral aspects of auditory signals. As the integration time windows needed for carrier frequency analysis would be much shorter than the windows needed for the analysis of the temporal pattern, "the N1m latency difference between spectral and temporal change conditions strongly suggests that the underlying neural mechanisms encoding spectral versus temporal information are different and require different temporal integration windows in the human auditory cortex" (Okamoto *et al.*, 2009: 6).

In a similar vein, Poeppel suggests that early stages of speech perception are mediated by both the left and the right hemisphere. Somehow this view is at odds with the notion of a left hemisphere dominance for speech that encompasses phonological, morphological, syntactic, and semantic computation. However, the computational mechanisms that are essential for the early prelexical stages of speech perception are evidently not restricted to the left hemisphere. Case studies in patients suffering from "pure word deafness" have demonstrated compellingly that severe distortion of speech perception only occurs when the cortical integrity of both the left and the right pARC is compromised. Based on this evidence Poeppel has developed a framework – the 'asymmetric sampling in time hypothesis' (AST) – in which functional asymmetries related to speech perception may be accounted for by different hemispheric preferences for temporal resolution: the left auditory

areas preferentially extract information over short temporal integration windows (~ 40 Hz, gamma band) and the right auditory areas over long integration windows (~ 4-10 Hz, theta and alpha bands) (Poeppel, 2001, 2003). In other words, temporal integration windows of different length should be considered the computational mechanism responsible for decoding the inflowing stream of auditory signals. As Poeppel (2001) maintains "acoustic-phonetic components such as formant transitions occur over short time scales, say 25-50 ms" (688). Complementary to Zatorre's model auditory information that constitutes suprasegmental events (e.g. intonation contour) are computed in long temporal integration windows (150-300 ms). According to Poeppel, auditory processing occurs symmetrically in the left and right primary auditory cortex which covers the medial two-third of Heschl's gyrus (Morosan, Rademacher, Schleicher *et al.*, 2001) during early stages of perception but the succeeding stages of analysis during which auditory input is recognized as speech are processed asymmetrically due to the temporal preferences of the cortical systems mentioned above.

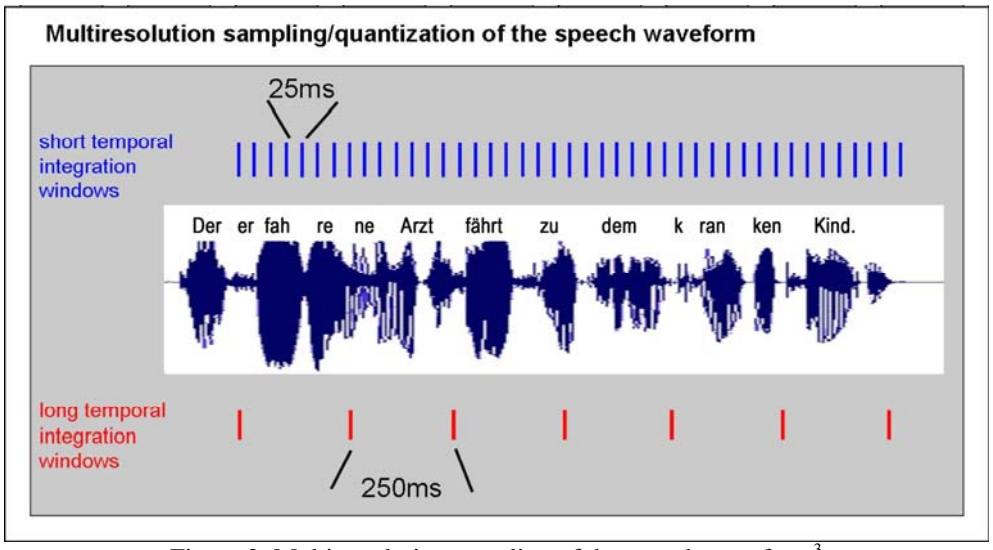


Figure 2: Multi resolution sampling of the speech waveform[3]

[3] The figure visualizes the waveform recorded from the German sentence *Der erfahrene Arzt fährt zu dem kranken Kind* ("The experienced doctor visits the sick child"). The blue bars represent short temporal integration windows at the ~ 40 Hz time range while the red bars indicate the long temporal integration windows at the ~ 4 Hz time range.

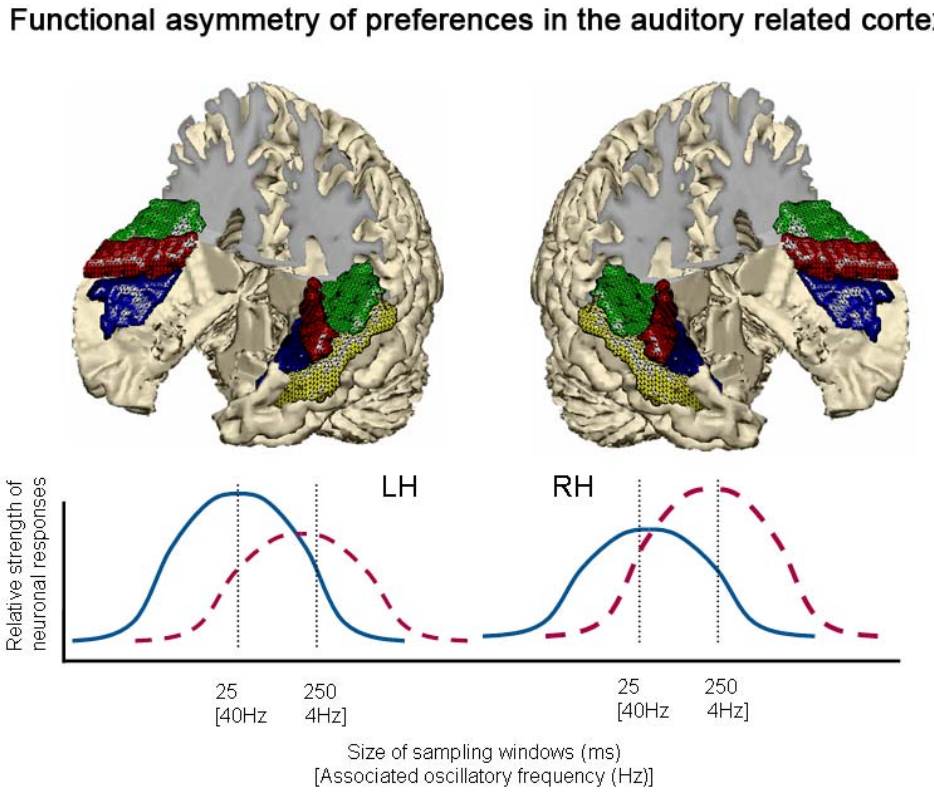


Figure 3: Functional asymmetry of computational preferences in the human auditory related cortex[4]

Due to the AST-hypothesis several predictions can be derived. Akin to non-speech sounds, speech signals enter the auditory cortex of both the left and right hemisphere. In a second stage the computation becomes asymmetric in that the left pARC preferentially computes (sub-)segmental information (i.e. formant transitions, rapid frequency modulated (FM) sweeps). Complementarily, the right pARC is more proficient at processing slowly changing, suprasegmental auditory information, namely aspects of prosody (speech melody and speech rhythm) but also features of music (instrumental timbre, melody) (Bever &

[4] The upper image shows the compartments that constitute the superior temporal region (blue = planum polare, red = transverse temporal gyrus, green = planum temporale, yellow = superior temporal sulcus). This ensemble of regions has been shown to accommodate the principal auditory functions. The lower illustration depicts the differential preferences of the two hemispheres with the left posterior auditory-related cortex being preferentially driven by rapidly changing cues and the right posterior auditory-related cortex being more amenable to slowly changing acoustic cues.

Chiarello, 1974). However, it should be emphasized that the model does not stipulate an exclusive attribution of the left pARC to segmental processing and an exclusive commitment of the right pARC to suprasegmental computation. The framework rather proposes a *preferential* processing mode and the time scale from 20 ms to 300 ms is regarded more as a continuum rather than a discrete scale which distinguishes sharply between 'short' and 'long' cues. One might now wonder to what extent the tenets of Poeppel and Zatorre should be considered complementary or alternative. With respect to the left hemisphere the answer is trivial. Both Zatorre's and Poeppel's model associate the left temporal cortex with the computation of transient phonetic features. Suprasegmental computation at the ~ 4 Hz range, however, has been associated with a particular sensitivity of the right pARC. Thus, due to Boemio, Fromm, Braun & Poeppel (2005) the right temporal mechanism would be ideal to form an "effective temporal unit for spectral analysis" (197). The two proposals are consistent in that they favor a division of labor between the auditory related cortex, each with a particular profile. However, the scheme forwarded by Poeppel bears one decisive advantage as it assumes a continuum of preferences ranging from rapid (LH) to slow (RH) modulations while Zatorre's model is less flexible in postulating a distinction between computation of temporal (LH) and spectral (RH) features.

2.2 Empirical evidence

During the past decade, a number of imaging studies on sublexical speech perception have been performed which provide evidence for the AST-hypothesis. A vintage of these studies will be introduced in the following section. We will begin by sketching recent research which has addressed the neural underpinnings of (sub-)segmental, local auditory processing at the prelexical level before introducing studies that have elucidated the neural mechanisms of suprasegmental, global auditory processing. Of course this compilation of studies does not pretend completeness but it covers research reports which have been cited quite frequently and thus could be considered sound evidence for the suggested division of labor between the two hemispheres in relation to speech perception.

One study that used parametrically varied non-speech stimuli observed bilateral responses to temporal acoustic modulations in auditory fields (Boemio *et al.*, 2005). Albeit this study failed to find clearly lateralized responses to either rapidly or slowly changing auditory patterns, it reports a hierarchical organization of elemental auditory processing. While the STG appears to be involved in the processing of temporal information regardless of frequency, the right STS is preferentially driven by slowly modulating signals. According to the authors, the data support the AST-hypothesis in which sounds are analyzed on two different time scales. However, the initial reasoning of the model is modified insofar as the results of Boemio and colleagues suggest a crucial role of the STS in addition to the supratemporal plane.

Jancke, Wustenberg, Scheich & Heinze (2002) demonstrated a particular sensitivity of the left and right PT for speech stimuli (consonant-vowel (CV) syllables compared to vowels, white noise, and tones) and stronger leftward responses to CV syllables (/da/, /ta/) than to vowels in the mid-STS and concluded that the processing of phonetic features like voice onset time and formant transitions is managed by the posterior auditory-related cortex. However, it should be mentioned that the authors did not observe unilateral left-sided responses but rather found bilateral posterior temporal regions to be preferentially driven by phonetic cues. In a follow-up fMRI study, Zaehle and colleagues sought to find

evidence for the assumption that the left auditory areas reported by Jancke and coworkers are generally driven by (sub-) segmental cues, namely voice onset times (VOT), regardless of whether they are tied up with linguistic content (Zaehle, Wustenberg, Meyer & Jancke, 2004). The authors used the same CV syllables as Jancke *et al.* (2002). The VOT of the syllables was manipulated to increase the difficulty of the CV task. (VOTs in ms for /da/ = 30, /ta/ = 40). A gap detection task was used to examine auditory temporal processing in non-speech stimuli. In more detail, the listener was presented with two streams of sounds, one of which had a brief silent period ('gap') at its temporal midpoint. The listener's task was to identify this signal and thus the shortest detectable gap ('gap threshold') was determined. Performing gap detection calls for a fine-grained analysis of rapidly changing acoustic cues. For this study, gap stimuli as used by Phillips, Taylor, Hall, Carr & Mossop (1997) were created; they consisted of two sound elements separated by a gap. The duration of the gaps was 8 ms and 32 ms (further details of the stimuli are described in the article by Zaehle *et al.*, 2004). The fMRI analysis revealed exclusively left-sided activations of primary and posterior association cortex during the perception of rapid temporal information. In particular, overlapping left supratemporal activation was evoked by both non-speech sounds and speech stimuli. Thus, these data clearly evidenced the existence of a neural circuit preferentially driven by rapid temporal information processing in the auditory domain.

This study should be considered an important contribution not only with respect to the paradigm used but also for its innovative methodological procedure. It was amongst the first fMRI studies that applied a 'silent' clustered acquisition protocol which allows auditory stimulus presentation devoid of detrimental scanner noise. This procedure opened new possibilities in the research field of auditory cognition and perception as it was now possible to study brain mechanisms that subserve subtle facets of auditory perception that are not contaminated by auditory cortex activation that is due to ambient scanner noise.[5]

Another fMRI study corroborated the role of the posterior auditory cortex for processing brief auditory events (Meyer, Zaehle, Gountouna *et al.*, 2005). In this study, participants performed an auditory AXB discrimination task on a set of sine-wave analogues that could be perceived as either non-speech or speech (Best, Morrongiello & Robson, 1981). In an uninformed condition participants listened naively to the stimuli. In the following break they were informed that the stimuli were sine-wave analogues of the spoken words *say* and *stay* that were synthesized from the recordings of a male native speaker. The sounds were described by naive listeners as unnatural synthetic sounds (*alien sounds*), while informed listeners had no difficulty in recognizing them as speech. To accomplish this perceptive shift participants have to focus their attention on the brief temporal gaps that help discriminate the speech sounds. Behavioral results revealed a difference in the processing mode; spectro-temporal integration occurred during (informed) speech perception, but not during uninformed condition while (naive) non-speech volunteers perceived only spectrally. The fMRI analyses yielded an activation increase in the adjacent portions of the left posterior auditory cortex (HG, PT, STS), suggesting an

[5] To learn more about 'silent fMRI' consult Gaab, Gabrieli & Glover (2007a, 2007b); Hall, Haggard, Akeroyd *et al.* (1999); Schmidt, Zaehle, Meyer *et al.* (2008); Zaehle, Schmidt, Meyer *et al.* (2007).

essential role of the pARC when the processing of rapidly changing auditory cues is emphasized.

Furthermore, results of a recent fMRI study provided evidence for the existence of two sublexical processing streams in the perisylvian cortex (Zaehle, Geiser, Alter, Jancke & Meyer, 2008). In this study the authors examined the auditory spectro-temporal processing in speech and non-speech sounds. Participants discriminated verbal and nonverbal auditory stimuli according to either spectral or temporal acoustic features. The results revealed specific activation in a dorsal stream involving the left IFG and the left parietal operculum when participants had to discriminate speech and non-speech stimuli based on subtle temporal acoustic features. By contrast, when participants had to discriminate the same stimuli based on changes in the frequency, bilateral activations along the middle temporal gyrus and STS lighted up. Thus, the results of this study demonstrate an involvement of the dorsal pathway in the segmental sublexical analysis of speech sounds as well as in the segmental acoustic analysis of non-speech sounds with analogous spectro-temporal characteristics and thus add to the present knowledge by suggesting two left perisylvian areas beyond the PT as being sensitive to subtle modulations occurring in the acoustic signal.

Depth EEG recording has also been employed to fathom the neural underpinnings of rapid auditory temporal processing (Liegeois-Chauvel, de Graaf, Laguitton & Chauvel, 1999; Trebuchon-Da Fonseca, Giraud, Badier, Chauvel & Liegeois-Chauvel, 2005). This approach has provided evidence that decoding of brief auditory features available in speech (CV syllables) and non-speech sounds mimicking the temporal structure of a syllable elicits neuronal activation in the left posterior auditory cortex. Due to the higher temporal sampling rate, the EEG technique is better suited than fMRI for investigating perception of temporal information at the range of milliseconds. However, in contrast to scalp EEG acquisition, depth EEG recording is more constrained as it is an invasive approach and does not allow a simultaneous estimation of neural sources from more remote brain sites. In order to corroborate the findings observed by intracranial EEG, Zaehle and colleagues performed a scalp EEG experiment in combination with a source estimation approach – LORETA – devoid of any assumptions about the location, number, and orientation of neuronal generators (Zaehle, Jancke & Meyer, 2007). In this study Zaehle and colleagues recorded and compared scalp auditory evoked potentials (AEP) in response to consonant-vowel syllables with varying VOT and non-speech analogues with varying noise-onset-time (NOT). Source estimation implied overlapping supratemporal networks involved in the perception of both speech and non-speech sounds with a bilateral activation pattern during the N1a time window and leftward asymmetry during the N1b time window. Elaborate regional statistical analysis of the activation over the middle and posterior portion of the STP revealed strong left lateralized responses over the middle STP for both the early and late component, and a functional leftward asymmetry over the posterior STP for the late component only. In congruency with the aforementioned recent brain imaging studies, the latter study supports the view that similar neural mechanisms underlie the perception of acoustically equivalent brief auditory events in speech and non-speech sounds.

In another recent scalp EEG study, Zaehle *et al.* (2009) used similar gap stimuli as in their aforementioned study of 2004, but this time they compared Mismatch-Negativity (MMN) responses evoked by either spectral or temporal deviants in order to find further evidence supporting the notion of an asymmetry of cortical tuning, with left and right

auditory areas being differentially sensitive to spectro-temporal features in acoustic signals and particularly in speech. By using MMN-responses in combination with LORETA, the authors were able to show a major pre-attentive contribution of the left perisylvian cortex to temporal deviant processing and of the right perisylvian cortex to the pre-attentive perception of spectral deviants. In particular, activity was found bilateral over the superior temporal regions, namely the superior temporal sulcus and superior temporal gyrus, with stronger left hemispheric MMN's to temporal deviants and stronger right hemispheric MMN's to spectral deviants, respectively. In addition, MMN latencies were larger for temporal than for spectral deviants, thus providing support for the already mentioned notion that a proper analysis of auditory signals requires different temporal integration windows in the human auditory cortex (Okamoto *et al.*, 2009).

According to the AST-hypothesis, the right pARC is most amenable to slow prosodic modulations in spoken language. Prosody is a component of the linguistic system and it describes acoustic phenomena such as word stress, sentence mode, and phrasing, which relate indirectly to the morpho-syntactic structure of utterances (Shattuck-Hufnagel & Turk, 1996). The term prosody is also used to refer to the acoustic correlates of these sound-based characteristics of spoken language, i.e. fundamental frequency, amplitude, and duration. Listeners can rely on these acoustic parameters when decoding the morpho-syntactic structure of the sentences to which they attend. Two fMRI studies (Meyer, Alter, Friederici, Lohmann & von Cramon, 2002; Meyer, Steinhauer, Alter, Friederici & von Cramon, 2004) examined the responsiveness of supratemporal areas to speech melody, i.e. the intonation contour available in spoken sentences. For this purpose, the authors degraded proper German sentences ("*Die hungrige Ärztin fährt zu dem kranken Kind*"), uttered by a female speaker, by applying the PURR-filtering procedure (for technical details of this filtering procedure see Meyer, Alter & Friederici, 2003; Sonntag & Portele, 1998). In order to degrade the segmental content the authors removed the spectral qualities of the speech signal up to the third harmonic. Thus, the signal derived from this filtering procedure consisted of the pure intonation contour, lacking syntactic and lexical information; it sounded like a blurred and unintelligible human voice heard from behind a door. Normal sentences and monotonous sounding sentences with flattened intonation, lacking dynamic pitch contours (for technical details of the artificial resynthesis yielding flattened speech see Meyer *et al.*, 2004), served as control stimuli. In line with the AST-framework, listening to pure prosodic information as compared to hearing propositional speech (with and without global intonational modulations) produced marked activation in the posterior segment of the right Sylvian fissure (PT, planum parietale) in two independent fMRI studies (Meyer *et al.*, 2002; Meyer *et al.*, 2004). Furthermore, the same contrast also elicited strong hemodynamic responses in the bilateral neostriatum (head of caudate, putamen), supporting recent clinical suggestions regarding the essential role of the basal ganglia in prosodic processing. According to Van Lancker Sidtis, Pachana, Cummings & Sidtis (2006), damage to the basal ganglia can result in an impaired appreciation of prosodic parameters (variability of fundamental frequency etc.) and may be accompanied by reduced efficacy in the use of prosody in communicative interactions (monopitch) in patients suffering from a dopamine insufficiency (Pell, Cheang & Leonard, 2006).

The issue of a rightward preference for slowly changing acoustic information was also addressed in another neuroimaging study performed by our research group. In this innovative study the neural correlates of rhythm processing in spoken sentences were

investigated (Geiser, Zaehle, Jancke & Meyer, 2008). Akin to intonation contour, rhythm is a pivotal structuring element of speech which is crucially involved in segmentation processes subserving speech perception. Perceptual processing of German pseudo-sentences spoken with an exaggerated (isochronous) or a conversational (non-isochronous) rhythm was compared with each other (for details of the stimulus corpus see Friederici, Meyer & von Cramon, 2000). The 'isochronously' spoken sentences followed a regular meter (i.e. iambs, trochees, dactyls) whereas the 'non-isochronously' spoken sentences followed an irregular meter (i.e. iambs or trochees with a dactyl interposed between two metrical feet). The volunteers had to perform either a rhythm task (explicit rhythm processing) or a prosody task (implicit rhythm processing) when listening to the sentences. Explicit processing of suprasegmental metrical patterns selectively revealed activation in the right posterior Sylvian segment (STP, IPL, parietal operculum). Implicit processing elicited activation in contralateral areas. These results strongly support the notion that the right posterior auditory cortex is involved in the explicit perception of auditory supra-segmental cues.

It has been argued that it is not suprasegmental but spectral information available in vocal stimuli that calls on right auditory engagement. Indeed, work of other groups (Beaucousin, Lacheret, Turbelin *et al.*, 2007; Belin & Zatorre, 2003; Belin, Zatorre & Ahad, 2002; Belin, Zatorre, Lafaille, Ahad & Pike, 2000; Chartrand & Belin, 2006; Fecteau, Armony, Joanette & Belin, 2004; Warren, Scott, Price & Griffiths, 2006) and our own work (Lattner, Meyer & Friederici, 2005; Meyer, Zysset, von Cramon & Alter, 2005) point to a general susceptibility of the entire right superior temporal region to voice spectral information and vocal timbre. However, if one undertakes a closer survey of the literature it becomes evident that these studies consistently report an involvement of anterior temporal lobe (STP/STS) circuits when it comes to decoding particular facets of human vocalization while the right pARC appears to be more susceptible to spectro-temporal modulations in vocal and nonvocal signals (Meyer, Baumann & Jancke, 2006; Meyer, Baumann, Wildgruber & Alter, 2007).

3. PRELIMINARY CONCLUSION

The preceding section summarized EEG and fMRI data collected by our group that supports the AST-hypothesis. The second part of this review explores the extent to which the available evidence corroborates the findings of other research groups that have addressed the same issue. This second part also points to structural asymmetries found at different neuroanatomical levels and which imply a specific endowment of the left pARC for rapid and efficient signal processing.

3.1 Corroborative evidence

During the last few years some reports have been published that also lend considerable credence to the view of an asymmetrical susceptibility of the pARC to differently modulated temporal acoustic information. Akin to our manifold evidence, there is an fMRI study in which overlapping hemodynamic responses were found in perisylvian territories, in particular in the left posterior STG/STS for processing rapid temporal cues in sublexical speech and non-speech signals (Joanisse & Gati, 2003). Interestingly, this study did not identify regions of activation that could be selectively attributed to speech when contrasted with non-speech signals. In line with this finding, a Norwegian group observed a functional

leftward asymmetry to the pARC triggered by rapid modulations during phonetic perception (Rimol, Specht, Weis, Savoy & Hugdahl, 2005). In keeping with Poeppel's proposal, another fMRI study in which parametrically varied non-speech stimuli were used observed bilateral responses to temporal modulations in auditory signals in auditory fields (Boemio *et al.*, 2005). Albeit this study failed to find clearly lateralized responses to either rapidly or slowly changing auditory patterns, rather a hierarchical organization of elemental auditory processing is reported. While the STG appears to be involved in the processing of temporal information regardless of frequency, the right STS is preferentially driven by slowly modulating signals. According to the authors, the data support the AST-model in which sounds are analyzed on two different time scales. However, the initial reasoning of the model is modified insofar as the results of Boemio and colleagues suggest a crucial role of the STS in addition to the supratemporal plane. Corroborative evidence has been provided by an fMRI study that assigned a critical role in phonetic perception to the STS (Benson, Richardson, Whalen & Lai, 2006). As a result, the STS has recently come under particular scrutiny with respect to speech perception and thus is considered part of the pARC. However, a recent fMRI study provided by Britton *et al.* (2009) examined the influence of spectral and durational properties on hemispheric asymmetries in vowel perception by applying speech (vowels) and non-speech (steady-state tones) stimuli, which varied in frequency and duration. Consistent with the described hypotheses about lateralization in auditory signal processing, there was a right hemisphere preference in the superior temporal gyrus for the processing of spectral information for both vowel and control (tone) stimuli. In particular, observed laterality differences for vowels and tones were "a function of heightened right hemisphere sensitivity to long integration windows, whereas the left hemisphere showed sensitivity to both long and short integration windows" (1). According to the authors, their findings challenge to some extent the strong version of the hypothesis that there are hemispheric differences in the integration window or time scale for processing spectral information (Hickock & Poeppel, 2004). The strong version would imply that short duration stimuli should show left lateralization, what actually has not been found in this study. Nevertheless, one should keep in mind that just by extending or shortening a certain stimulus, lateralization has not to be affected necessarily, as it has not yet been established whether such a procedure changes the fundamental processing demands of the particular stimulus by any means. Furthermore, as activation increased bilaterally as a function of stimulus duration in similar ways, it is rather unlikely such a change occurred.

In an innovative MEG study, Luo & Poeppel (2007) convincingly showed that the prominent phase pattern of theta band (4-8Hz) responses corresponds to the ability to discriminate spoken sentences. According to such recent evidence, inflowing spoken utterances are segmented in an ~ 200 ms time window (period of theta oscillation) and subjected to further analysis in the pARC, once typical speech dynamics have been identified. As speech prosody is a dynamic feature of speech it comes as no surprise that the right pARC responds more preferentially to it.

There is some evidence which has challenged the proposed right pARC preference for slow acoustic modulations typically representing intonation contour in spoken utterances. For example, Hesling and coworkers observed that sensory prosodic integration preferentially triggered right posterior periauditory sites but also elicited activations in right and left inferior frontal and extrasylvian areas (Hesling, Clement, Bordessoules & Allard,

2005; Hesling, Dilharreguy, Clement, Bordessoules & Allard, 2005). However, this finding fits in well with other reasoning about the hemispheric roles in the perception of prosody. Recent work has compellingly demonstrated that the functional asymmetry of speech prosody in frontoopercular areas can be influenced appreciably by individual language experience, language lateralization and specific task demands (Gandour, Dzemidzic, Wong *et al.*, 2003; Gandour, Tong, Talavage *et al.*, 2007; Gandour, Tong, Wong *et al.*, 2004).

3.2 Neuroanatomical considerations

Complementary to the aforementioned studies a handful of observations have been published that addressed the issue of structural asymmetry of the perisylvian cortex. Even though anatomical brain asymmetry is found in all mammals, its link to functional lateralization is still elusive. At least the relationship between structural asymmetry and language lateralization has been extensively investigated and could be considered established knowledge. It is interesting to note that the gross anatomical asymmetry of the PT is already present around the 30th week of gestation (Steinmetz, 1996) and might be triggered externally by asymmetric exposure to auditory in-utero stimulation (Toga & Thompson, 2003). Albeit Galaburda and colleagues concluded that the volumetric leftward asymmetry of the PT should be considered a significant marker of leftward speech lateralization (Galaburda *et al.*, 1978), subsequent neuroanatomical work has demonstrated that this asymmetry is most prominent in absolute pitch possessors (Schlaug, Jancke, Huang & Steinmetz, 1995). This finding convincingly suggests that the planum temporale plays a pivotal role in auditory processing in general rather than being selectively tied to language because the absolute pitch possessors' acuity has developed as a function of musical training and cannot be related to a particular speech proficiency.

More recently, the structural leftward asymmetry of posterior auditory-related regions supporting the notion of a LH preference for processing fine-grained elemental auditory features has been evidenced at various macro- and microscopic levels. The posterior end of the Sylvian fissure is higher in the RH than in the LH in nearly 70% of right-handers – a fact taken to be indirect evidence for leftward lateralization of speech functions (LeMay, 1982). Similarly, the STS extends 7-9 mm further back in the LH which is suggestive of a larger surface area of the left posterior auditory cortex (Sowell, Thompson, Rex *et al.*, 2002). Furthermore, the right STS is significantly deeper than the left STS while the small strips of fiber bundles bridging the fund of the STS and connecting the auditory cortex and the inferior temporo-occipital association areas are more prominent on the LH (Ochiai, Grimault, Scavarda *et al.*, 2004). By means of MR-based morphometry, volumetry and automatic voxel-based morphometry (for detailed explanation of the latter observer-independent approach see Ashburner & Friston, 2000, 2001) recent studies demonstrated a strong volumetric leftward asymmetry of the HG/PT complex (Knaus, Bollich, Corey, Lemen & Foundas, 2006) and a significant leftward grey matter PT asymmetry (Dorsaint-Pierre, Penhune, Watkins *et al.*, 2006). A postmortem study in sixteen specimens revealed greater white matter volume in the left posterior temporal lobe (Anderson, Southern & Powers, 1999). Akin to this finding Sigalovsky, Fischl & Melcher (2006) reported greater grey matter myelination in left primary and secondary auditory cortex. Due to these authors this finding is suggestive of a "substrate for the left hemisphere's specialized processing of speech, language, and rapid acoustic changes" (1524). Interestingly, deaf individuals who grew up without exposure to spoken language have been found to display less prominent white matter tracts in the left auditory cortex and the left posterior arcuate fascicle (AF)

compared to hearing controls, indicating a relationship between the degree of myelination of the left pARC and lifelong exposure to speech and non-speech sounds (Emmorey, Allen, Bruss, Schenker & Damasio, 2003; Meyer, Toepel, Keller *et al.*, 2007). Furthermore, recent diffusion tensor tractography studies have demonstrated greater white matter morphology and more pronounced fiber density in the left arcuate fascicule (Barrick, Lawes, Mackay & Clark, 2007; Nucifora, Verma, Melhem, Gur & Gur, 2005). Remarkably, this leftward asymmetry of the arcuate fascicle occurs regardless of handedness or language lateralization (Vernooij, Smits, Wielopolski *et al.*, 2007) which prompted the authors to suggest that this asymmetry could be driven by the involvement of the arcuate fascicle in other tasks such as elemental acoustic processing rather than language processing per se. In essence, these findings suggest greater grey matter myelination in the left posterior auditory-related cortex, which might be a neuroanatomical substrate for the LH's specific preference for speech and rapid acoustic changes, because more pronounced myelination would facilitate faster and more efficient stimulus processing (Chiarello, Kacinik, Manowitz, Otto & Leonard, 2004). In a similar vein, Glasser and Rilling used DTI to track arcuate fascicle connections between the cortical regions implicated in various aspects of speech processing and assessed their degree of laterality. Glasser and Rilling compared their DTI tractography results with activation coordinates from prior neuroimaging studies having researched phonologic, lexical-semantic and prosodic processing. Using this procedure, the authors were able to identify two distinct pathways within the AF, one linking the posterior part of the STG to the frontal lobe and the other connecting the MTG to the frontal lobe. Remarkably, these pathways were not found in all of the subjects participating in this study. Mainly in the right hemisphere, only a few subjects had an STG pathway (4/20), but also the right hemispheric MTG pathway was absent in almost half of the study's participants (11/20). In the left hemisphere, the MTG-frontal lobe connection was found in all subjects (20/20) and nearly all subjects exhibited a left hemispheric STG pathway (17/20). According to the authors, the STG connection is a much smaller pathway than the MTG and therefore it may be more vulnerable to crossing fibers or motion artifacts, which may have prevented its identification in the left hemisphere of some subjects. However, regarding the absolute difference between left and right volumes, significant leftward asymmetries were found in both AF segments, with the STG pathway being more asymmetric when considering the number of subjects in which a right hemisphere pathway was present. Furthermore, functional activations from studies investigating phonologic processing were bilateral and overlapped with the termination sites of the STG pathway in the left hemisphere. In the right hemisphere, the foci of phonologic activations were located more anterior in the STG/STS and did not overlap with the STG pathway in those subjects in which this pathway was detectable. Nevertheless, activations in studies investigating prosodic processing overlapped with both the MTG and the STG segments within the right hemisphere, "suggesting that in some subjects, the cortex involved in prosodic processing extends beyond the superior bank of the STS and into a small portion of the posterior STG" (2474). On the other hand, activations from lexical-semantic tasks were found throughout the middle and inferior temporal gyri along with the angular gyrus of the left hemisphere and were concentrated over the termination sites of the left hemispheric MTG pathway. So in sum, for the left hemisphere, temporal lobe activations in phonologic processing tasks overlapped with the termination of the STG pathway, whereas lexical-semantic activations overlapped with the termination of the MTG

pathway. In contrary, in those subjects who had a right hemispheric STG connection, it did not overlap with phonologic activations, which were located more anterior in the right mid-STG/STS than they are in the left, whereas prosodic activations fitted both the right hemispheric STG and MTG connection sites. Taken together, these findings show that the pattern of pathway asymmetry does not always correlate with lateralization of function inferred from functional neuroimaging studies. For example, while the STG is suggested to be bilaterally involved in phonologic processing (e.g. Jancke *et al.*, 2002; Specht *et al.*, 2003), only the left hemisphere has a strong and consistent connection to the frontal lobe via the AF. Anyhow, Glasser and Rilling reasonably interpreted their results within the framework of the language model of Price (2000) and Hickok & Poeppel (2004). Briefly summarized, auditory information is first processed bilateral in primary auditory related cortex and is then decoded phonologically in the left hemispheric posterior BA 22. From there it can be conveyed directly via the STG pathway on to Broca's area if it's to be repeated immediately, or it can be relayed to the cortex below the STS for lexical-semantic analysis. Information can then be exchanged between temporal and frontal lobes through the bi-directional connections that comprise the MTG pathway in order to establish high level lexical-semantic processing required to form logical and coherent speech. Thus, according to the authors, "the direct phonological loop connecting the posterior STG to Broca's area might be particularly important during the acquisition of language by children, as it allows decoded phonemes direct access to Broca's area for speech output, whereas the MTG pathway might be most important for carrying lexical-semantic information during spontaneous production of established speech" (Glasser & Rilling, 2008: 2475). In the right hemisphere, the sole fronto-temporal connection consistently found links the posterior MTG and the frontal lobe, which is the area that is activated in prosodic compared to linguistic processing (e.g. Meyer *et al.*, 2002; Riecker *et al.*, 2002), even though this link is much weaker than the homologous left hemispheric pathway. For those subjects who had an STG pathway in the right hemisphere, activation patterns due to prosodic processing also lined up with STG's posterior termination area, thus suggesting "that the STG and MTG segments may not have distinct functions in the right hemisphere" (Glasser & Rilling, 2008: 2475). Taken together, these findings provide anatomical evidence for the phonologic and lexical-semantic pathways postulated by the model of Price (2000) and Hickok & Poeppel (2004) as well as for a right hemispheric pathway between the MTG and the frontal lobe that is hypothesized to be involved in prosodic processing (Ethofer *et al.*, 2006).

Recently, Friederici (2009) reviewed several diffusion tensor imaging (DTI) studies in order to shed some further light on the white matter fiber tracts connecting circumscribed brain regions associated with speech processing. In essence, Friederici focused on fiber tracts connecting prominent language-relevant areas in the left hemisphere, namely subdivisions of the inferior frontal gyrus (IFG) and several temporal lobe structures including superior temporal gyrus (STG), superior temporal sulcus (STS) and middle temporal gyrus (MTG). In addition to the already described AF connecting perisylvian areas, DTI studies indicate that the AF is not the only white matter tract connecting language relevant regions. One additional "dorsal" pathway connects BA 44 via the superior longitudinal fasciculus (SLF) to the posterior temporal lobe, in particular the lateral STG and MTG. Furthermore, this pathway also exhibits connectivities to BA 40 located in the inferior parietal lobe. Two more ventrally located routes run from anterior Broca's area (BA 45) through the ventral

portion of the extreme capsule (ECFS) and the uncinate fascicle (UF) to the anterior STG. As the current DTI methods are limited in resolution, the two dorsal (the AF and SLF) and ventral (ECFS and UF) pathways are not reliably separable because of their neuroanatomical adjacency and therefore must await further empirical support. Despite these limitations, two of these pathways were defined as being relevant for syntactic processes. First, one ventral pathway connects the frontal operculum to the anterior STG via the UF, two structures which have been shown to be involved in phrase structure building (Friederici *et al.*, 2003; 2006). Second, two regions which are known to subserve the processing of syntactically complex sentences, in particular the pars opercularis and the posterior portion of the STG/STS (Bornkessel *et al.*, 2005), are linked dorsally via the SLF. On the other hand, the ventral ECFS pathway connecting the pars triangularis and orbitalis to the mid portion of the STG/MTG appears to support semantic processes in the adult human brain, as it has been shown that the pars triangularis and the pars orbitalis are involved in the processing of semantic information (Vigneau *et al.*, 2006). Other possible functions of the dorsal and ventral routes have been indicated in a recent combined functional and DTI study by Saur *et al.* (2008). They have found by using language comprehension and sublexical repetition tasks that the dorsal route might be responsible for sound-to-meaning mapping and the ventral route via the ECFS for aspects of language comprehension. So taken together, the data reviewed by Friederici (2009) indicate "that there are several pathways connecting the language-relevant brain areas with the dorsal pathway connecting the posterior part of Broca's Area (BA 44) and the posterior STG/STS being crucial for the human language capacity, which is characterized by the ability to process complex sentence structures". This assumption receives further support by the finding that non-human primates not being capable of learning and processing hierarchically structured sequences (Fitch & Hauser, 2004) do not seem to possess a strong dorsal connection between BA 44 and the posterior STG/STS (Schmahmann *et al.*, 2007; Rilling *et al.*, 2008). Furthermore, the dorsal pathway connecting the language areas is not fully myelinized in children at an age in which they are still deficient in processing syntactically complex sentences (Friederici, 2009). Thus, "it seems that the evolution of language is tightly related to the maturation of the dorsal pathway connecting those areas which in the adult human brain are involved in the processing of syntactically complex sentences" (Friederici, 2009: 180). In corroboration of this argument, we point to another recently published study demonstrating that the dorsal pathway matures last amongst all language-related brain regions (Su *et al.*, 2008).

At the microscopic level, there exists also evidence that the two hemispheres differ with respect to columnar architecture in the auditory cortex (Hutsler & Galuske, 2003). In particular, the spacing of distinct cortical micro-columns in the left auditory cortex is wider which accounts for automatic and rapid signal processing (Galuske, Schlote, Bratzke & Singer, 2000; Hutsler, 2003).

However, one should notice that macroanatomical leftward asymmetries of the posterior Sylvian fissure, especially of the PT have also been described in great apes (Cantalupo, Pilcher & Hopkins, 2003; Gannon, Holloway, Broadfield & Braun, 1998) that do not share the faculty of language with the human species. This finding deserves a short remark. Should the conjecture hold true that leftward asymmetry in the brain of humans and great apes may be associated with a superior endowment of the left pARC for processing subtle acoustic information, it becomes less perspicuous to directly associate this asymmetry with

lateralization of language functions. Furthermore, Steinschneider, Volkov, Fishman *et al.* (2005) have demonstrated by means of electrophysiological recordings that monkeys are also able to discriminate brief variable tone onset times at the range of 20 ms. Based on this observation it is questionable to assume that the ability to process acoustic signals with fine grained resolution are the sole computational basis for speech perception.

On the other hand, gross analyses of large scale cerebral areas, as mentioned above, may ignore small but existing structural differences between species. By studying Nissl-stained slides of normal human, chimpanzee, and rhesus monkey brains in a region of the PT, a significant leftward asymmetry only in the human brain was unveiled (Buxhoeveden, Switala, Litaker, Roy & Casanova, 2001). The human PT showed wider columns and more neuropil space in the left hemisphere while this asymmetry could not be found in chimpanzee and rhesus monkey brains. Future research will need to explore what this difference might mean and to what extent it may constitute the cortical basis of phoneme detection. However, for the time being and based on the present knowledge it should be allowed to assume that such a basis exists.

3.3 Final remarks

The present review introduced a series of neuroimaging studies that explored the neural underpinnings of early stages of speech perception. In agreement with the AST-hypothesis (Poeppel, 2001, 2003) the presented findings supply evidence for a 'division of labor' between the left and right core auditory (HG) and adjoining auditory-related cortical areas (PT, STS, planum parietale, parietal operculum). The initial version of the model suggested that auditory processing occurs symmetrically in the core area bilaterally and occurs asymmetrically in the auditory-related cortical areas. However, the confluence of the aforementioned neuroimaging studies in which the framework's predictions were examined have observed functional asymmetry in all posterior perisylvian regions including the auditory core areas. MR data on structural asymmetry in these brain regions corroborate the functional observations. Furthermore, it presently seems that the left pARC is basically equally amenable to rapidly and slowly changing acoustic cues while a preference of the right pARC to slow modulations of the acoustic signal has been clearly demonstrated (Poeppel, Idsardi & van Wassenhove, 2008).

Most important, it should be mentioned that the present review emphasizes the early stages of speech perception. In other words, higher levels of language processing incorporating lexical, syntactic, or semantic information and recruiting cortical areas beyond the auditory-related cortex are not a primary subject of this review. Models addressing the issue of brain mechanisms underlying higher levels of language processing have been forwarded (Friederici, 2002; Friederici & Alter, 2004; Grodzinsky & Friederici, 2006; Hickok & Poeppel, 2000, 2004, 2007). While the latter account is limited to the word-level, the theoretical blueprints of Friederici as well as of Hickok & Poeppel examine how the phonological information is integrated with knowledge about syntax and semantics implemented in the human brain. All these approaches emphasize the eminent role of the left anterior perisylvian cortex as principal site of higher language functions. Thus, the present evidence consistently indicates the existence of a general speech processing stream in the human brain (Hickok & Poeppel, 2007; Scott & Wise, 2004). While elemental acoustic, phonetic and prosodic perception appears to recruit primarily the posterior auditory-related cortical areas, the left anterior STP, the left anterior STS and the left frontal operculum are associated with higher facets of language (Friederici *et al.*, 2006;

Heinke, Fiebach, Schwarzbauer *et al.*, 2004; Humphries, Binder, Medler & Liebenthal, 2007; Humphries, Love, Swinney & Hickok, 2005; Meyer *et al.*, 2003, 2004).

The present paper further underlines the importance of suprasegmental acoustic cues, e.g. speech rhythm and speech melody as essential structural elements that help the listener group words and phrases to perform a more efficient integration of syntactic and semantic information and to achieve a proper representation of a spoken utterances . In particular, the interplay between large scale motor-related and auditory-related networks during the perception of speech rhythm and speech melody underscores the elemental meaning of these mechanisms for principal facets of human behavior, namely action and language (Willems & Hagoort, 2007). In essence, the review points to an array of recent imaging studies that have elucidated the neural underpinnings of sublexical, subsegmental and suprasegmental speech perception. In accordance with the parameter-based AST-hypothesis (Poeppel, 2001, 2003), our data clearly demonstrate a substantial involvement of the right posterior perisylvian cortex in processing suprasegmental cues (speech melody, speech rhythm) as well as vocal and instrumental timbre. Furthermore the data suggest a preference of the left posterior perisylvian cortex for processing subsegmental, rapidly changing acoustic features. Thus, in concordance with Poeppel & Embick (2005), there is a strong need for an extensive revision of the classical 19th and 20th neurological models of language processing that have emphasized the brain mechanisms of linguistic components, namely phonology, syntax, and semantics, more than substantiating the importance of (supra)segmental information during speech perception. Moreover, recent studies investigating speech processing in cochlear implant users (Sandmann *et al.*, 2009) and absolute pitch possessors (Oechslin *et al.*, 2009) have shown the existence of alternative auditory processing routines deviating to some extent from the ones described in this article. Therefore, further research regarding various aspects of fundamental speech processing has to be done not only with normal-hearing subjects, but also with special populations such as cochlear-implant users or absolute pitch possessors, in order to achieve a broad understanding of the human language ability.

ACKNOWLEDGMENTS

The authors are grateful to David Poeppel for providing illustrations included in figures of this article. Current research on the neurocognitive mechanisms underpinning speech perception is supported by *Schweizerischer Nationalfonds* (account number 320000-120661-1).

4. REFERENCES

Amunts, K. & Zilles, K. (2006), A multimodal analysis of structure and function in Broca's region. In *Broca's region* (Y. Grodzinsky & K. Amunts, editors), New York: Oxford University Press, 17-30.

Anderson, B., Southern, B. D. & Powers, R. E. (1999), Anatomic asymmetries of the posterior superior temporal lobes: a postmortem study, *Neuropsychiatry Neuropsychol Behav Neurol*, *12*, 247-254.

Ashburner, J. & Friston, K. J. (2000), Voxel-based morphometry – the methods, *Neuroimage, 11*, 805-821.

Ashburner, J. & Friston, K. J. (2001), Why voxel-based morphometry should be used, *Neuroimage, 14*, 1238-1243.

Barrick, T. R., Lawes, I. N., Mackay, C. E. & Clark, C. A. (2007), White matter pathway asymmetry underlies functional lateralization, *Cereb Cortex, 17*, 591-598.

Beaucousin, V., Lacheret, A., Turbelin, M. R., Morel, M., Mazoyer, B. & Tzourio-Mazoyer, N. (2007), FMRI study of emotional speech comprehension, *Cereb Cortex, 17*, 339-352.

Belin, P. & Zatorre, R. J. (2003), Adaptation to speaker's voice in right anterior temporal lobe, *Neuroreport, 14*, 2105-2109.

Belin, P., Zatorre, R. J. & Ahad, P. (2002), Human temporal-lobe response to vocal sounds, *Brain Res Cogn Brain Res, 13*, 17-26.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P. & Pike, B. (2000), Voice-selective areas in human auditory cortex, *Nature, 403*, 309-312.

Belin, P., Zilbovicius, M., Crozier, S., Thivard, L., Fontaine, A., Masure, M. C. & Samson, Y. (1998), Lateralization of speech and auditory temporal processing, *J Cogn Neurosci, 10*, 536-540.

Ben Shalom, D. & Poeppel, D. (2008), Functional anatomic models of language: assembling the pieces, *Neuroscientist, 14*, 119-127.

Benson, R. R., Richardson, M., Whalen, D. H. & Lai, S. (2006), Phonetic processing areas revealed by sinewave speech and acoustically similar non-speech, *Neuroimage, 31*, 342-353.

Best, C. T., Morrongiello, B. & Robson, R. (1981), Perceptual equivalence of acoustic cues in speech and non-speech perception, *Percept Psychophys, 29*, 191-211.

Bever, T. G. & Chiarello, R. J. (1974), Cerebral dominance in musicians and nonmusicians, *Science, 185*, 537-539.

Binkofski, F. & Buccino, G. (2004), Motor functions of the Broca's region, *Brain Lang, 89*, 362-369.

Birbaumer, N. & Schmidt, R. F. (2002), *Biologische Psychologie*, Berlin: Springer.

Boemio, A., Fromm, S., Braun, A. & Poeppel, D. (2005), Hierarchical and asymmetric temporal sensitivity in human auditory cortices, *Nat Neurosci, 8*, 389-395.

Bogen, J. E. & Bogen, G. M. (1976), Wernicke's region – Where is it? *Ann N Y Acad Sci, 280*, 834-843.

Bornkessel, I., Zysset, S., Friederici, A. D., von Cramon, D. Y. & Schlesewsky, M. (2005), Who did what to whom? The neural basis of argument hierarchies during language comprehension, *Neuroimage, 26*, 221-233.

Brett, M., Johnsrude, I. S. & Owen, A. M. (2002), The problem of functional localization in the human brain, *Nat Rev Neurosci, 3*, 243-249.

Britton, B., Blumstein, S. E., Myers, E. B. & Grinrod, C. (2009), The role of spectral and durational properties on hemispheric asymmetries in vowel perception, *Neuropsychologia,* 47, 1096-1106.

Broca, P. (1863), Localisation des fonctions cérébrales: siège de langage articulé, *Bulletin de la Société d'Anthropologie de Paris,* 4, 200-208.

Buxhoeveden, D. P., Switala, A. E., Litaker, M., Roy, E. & Casanova, M. F. (2001), Lateralization of minicolums in human planum temporale is absent in nonhuman primate cortex, *Brain, Behavior and Evolution,* 57*,* 349-358.

Cantalupo, C., Pilcher, D. L. & Hopkins, W. D. (2003), Are planum temporale and sylvian fissure asymmetries directly related? A MRI study in great apes, *Neuropsychologia,* 41, 1975-1981.

Chartrand, J. P. & Belin, P. (2006), Superior voice timbre processing in musicians, *Neurosci Lett,* 405, 164-167.

Chiarello, C., Kacinik, N., Manowitz, B., Otto, R. & Leonard, C. (2004), Cerebral asymmetries for language: evidence for structural-behavioral correlations, *Neuro-psychology,* 18, 219-231.

Crinion, J. & Price, C. J. (2005), Right anterior superior temporal activation predicts auditory sentence comprehension following aphasic stroke, *Brain, 128*, 2858-2871.

Dapretto, M. & Bookheimer, S. Y. (1999), Form and content: dissociating syntax and semantics in sentence comprehension, *Neuron,* 24, 427-432.

Demonet, J. F., Thierry, G. & Cardebat, D. (2005), Renewal of the neurophysiology of language: functional neuroimaging, *Physiol Rev,* 85, 49-95.

Dorsaint-Pierre, R., Penhune, V. B., Watkins, K. E., Neelin, P., Lerch, J. P., Bouffard, M. & Zatorre, R. J. (2006), Asymmetries of the planum temporale and Heschl's gyrus: relationship to language lateralization, *Brain,* 129, 1164-1176.

Efron, R. (1963), Temporal Perception, Aphasia and Déjà Vu, *Brain, 86*, 403-424.

Embick, D., Marantz, A., Miyashita, Y., O'Neil, W. & Sakai, K. L. (2000), A syntactic specialization for Broca's area. *Proc Natl Acad Sci USA,* 97, 6150-6154.

Emmorey, K., Allen, J. S., Bruss, J., Schenker, N. & Damasio, H. (2003), A morphometric analysis of auditory brain regions in congenitally deaf adults, *Proc Natl Acad Sci USA,* 100, 10049-10054.

Ethofer, T., Anders, S., Erb, M., Herbert, C., Wiethoff, S., Kissler, J., Grodd, W. & Wildgruber, D. (2006), Cerebral pathways in processing of affective prosody: a dynamic causal modelling study, *Neuroimage,* 30, 580-587.

Fadiga, L., Craighero, L. & Roy, A. (2006), Broca's region: a speech area? In *Broca's region* (Y. Grodzinsky & K. Amunts, editors), New York: Oxford University Press, 137-152.

Fecteau, S., Armony, J. L., Joanette, Y. & Belin, P. (2004), Is voice processing species-specific in human auditory cortex? An fMRI study, *Neuroimage*, 23, 840-848.

Fink, G. R., Manjaly, Z. M., Stephan, K. E., Gurd, J. M., Zilles, K., Amunts, K. & Marshall, J. C. (2006), A role for Broca's area beyond language processing: evidence from neuropsychology and fMRI. In *Broca's region* (Y. Grodzinsky & K. Amunts, editors), New York: Oxford University Press, 254-268.

Fitch, W. T. & Hauser, M. D. (2004), Computational constraints on syntactic processing in a nonhuman primate, *Science,* 303, 377-380.

Friederici, A. D. (2009), Pathways to language: fiber tracts in the human brain, *Trends Cogn Sci.,* 13(4), 175-81.

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., Anwander, A. (2006), The brain differentiates human grammars: functional localization and structural connectivity, *Proc. Natl Acad Sci USA,* 103, 2458-2463.

Friederici, A. D., Rüschemeyer, S. A,, Hahne, A,, Fiebach, C. J. (2003), The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes, *Cereb Cortex,* 13, 170-177.

Friederici, A. D. (2002), Towards a neural basis of auditory sentence processing, *Trends Cogn Sci,* 6, 78-84.

Friederici, A. D. (2004), Processing local transitions versus long-distance syntactic hierarchies, *Trends Cogn Sci,* 8, 245-247.

Friederici, A. D. (2006), Broca's area and the ventral premotor cortex in language: functional differentiation and specificity, *Cortex*, 42, 472-475.

Friederici, A. D. & Alter, K. (2004), Lateralization of auditory language functions: a dynamic dual pathway model, *Brain Lang,* 89, 267-276.

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I. & Anwander, A. (2006), The brain differentiates human and non-human grammars: functional localization and structural connectivity, *Proc Natl Acad Sci USA,* 103, 2458-2463.

Friederici, A. D., Meyer, M. & von Cramon, D. Y. (2000), Auditory language comprehension: an event-related fMRI study on the processing of syntactic and lexical information, *Brain Lang,* 74, 289-300.

Gaab, N., Gabrieli, J. D. & Glover, G. H. (2007a), Assessing the influence of scanner background noise on auditory processing. I. An fMRI study comparing three experimental designs with varying degrees of scanner noise, *Hum Brain Mapp,* 28, 703-720.

Gaab, N., Gabrieli, J. D. & Glover, G. H. (2007b), Assessing the influence of scanner background noise on auditory processing. II. An fMRI study comparing auditory processing in the absence and presence of recorded scanner noise using a sparse design, *Hum Brain Mapp*, 28, 721-732.

Galaburda, A. M., Sanides, F. & Geschwind, N. (1978), Human brain. Cytoarchitectonic left-right asymmetries in the temporal speech region. *Arch Neurol,* 35, 812-817.

Galuske, R. A., Schlote, W., Bratzke, H. & Singer, W. (2000), Interhemispheric asymmetries of the modular structure in human temporal cortex, *Science,* 289, 1946-1949.

Gandour, J., Dzemidzic, M., Wong, D., Lowe, M., Tong, Y., Hsieh, L., Satthamnuwong, N. & Lurito, J. (2003), Temporal integration of speech prosody is shaped by language experience: an fMRI study, *Brain Lang,* 84, 318-336.

Gandour, J., Tong, Y., Talavage, T., Wong, D., Dzemidzic, M., Xu, Y., Li, X. & Lowe, M. (2007), Neural basis of first and second language processing of sentence-level linguistic prosody, *Hum Brain Mapp*, 28, 94-108.

Gandour, J., Tong, Y., Wong, D., Talavage, T., Dzemidzic, M., Xu, Y., Li, X. & Lowe, M. (2004), Hemispheric roles in the perception of speech prosody, *Neuroimage*, 23, 344-357.

Gannon, P. J., Holloway, R. L., Broadfield, D. C. & Braun, A. R. (1998), Asymmetry of chimpanzee planum temporale: humanlike pattern of Wernicke's brain language area homolog, *Science*, 279, 220-222.

Geiser, E., Zaehle, T., Jancke, L. & Meyer, M. (2008), The neural correlate of speech rhythm as evidenced by metrical speech processing: an fMRI study, *Journal of Cognitive Neuroscience,* 20, 541-552.

Geschwind, N. (1979), Specializations of the human brain. *Sci Am,* 241, 180-199.

Glasser, M. F. & Rilling, J. K. (2008), DTI Tractography of the Human Brain's Language Pathways, *Cerebral Cortex,* 18, 2471-2482

Griffiths, T. D. & Warren, J. D. (2002), The planum temporale as a computational hub, *Trends Neurosci,* 25, 348-353.

Grodzinsky, Y. (2000), The neurology of syntax: language use without Broca's area, *Behav Brain Sci,* 23, 1-21; discussion 21-71.

Grodzinsky, Y. (2006), The language faculty, Broca's region, and the mirror system. *Cortex,* 42, 464-468.

Grodzinsky, Y. & Friederici, A. D. (2006), Neuroimaging of syntax and syntactic processing, *Curr Opin Neurobiol,* 16, 240-246.

Hackett, T. A. & Kaas, J. H. (2004), Auditory cortex in primates: functional subdivisions and processing streams. In *The New Cognitive Neuroscience* (M.A. Gazzaniga, editor), 3rd ed. Cambridge, MA: MIT Press, 215-232.

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., Gurney, E. M. & Bowtell, R. W. (1999), 'Sparse' temporal sampling in auditory fMRI, *Hum Brain Mapp,* 7, 213-223.

Heinke, W., Fiebach, C. J., Schwarzbauer, C., Meyer, M., Olthoff, D. & Alter, K. (2004), Sequential effects of propofol on functional brain activation induced by auditory language processing: an event-related functional magnetic resonance imaging study, *Br J Anaesth,* 92, 641-650.

Hesling, I., Clement, S., Bordessoules, M. & Allard, M. (2005), Cerebral mechanisms of prosodic integration: evidence from connected speech, *Neuroimage,* 24, 937-947.

Hesling, I., Dilharreguy, B., Clement, S., Bordessoules, M. & Allard, M. (2005), Cerebral mechanisms of prosodic sensory integration using low-frequency bands of connected speech, *Hum Brain Mapp,* 26, 157-169.

Hickok, G. & Poeppel, D. (2000), Towards a functional neuroanatomy of speech perception, *Trends Cogn Sci,* 4, 131-138.

Hickok, G. & Poeppel, D. (2004), Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language, *Cognition*, 92, 67-99.

Hickok, G. & Poeppel, D. (2007), The cortical organization of speech processing, *Nat Rev Neurosci*, 8, 393-402.

Hoen, M., Pachot-Clouard, M., Segebarth, C. & Dominey, P. F. (2006), When Broca experiences the Janus syndrome: an ER-fMRI study comparing sentence comprehension and cognitive sequence processing, *Cortex,* 42, 605-623.

Humphries, C., Binder, J. R., Medler, D. A. & Liebenthal, E. (2007), Time course of semantic processes during sentence comprehension: an fMRI study, *Neuroimage,* 36, 924-932.

Humphries, C., Love, T., Swinney, D. & Hickok, G. (2005), Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing, *Hum Brain Mapp,* 26, 128-138.

Hutsler, J. & Galuske, R. A. (2003), Hemispheric asymmetries in cerebral cortical networks, *Trends Neurosci,* 26, 429-435.

Hutsler, J. J. (2003), The specialized structure of human language cortex: pyramidal cell size asymmetries within auditory and language-associated regions of the temporal lobes, *Brain Lang,* 86, 226-242.

Jamison, H. L., Watkins, K. E., Bishop, D. V. & Matthews, P. M. (2006), Hemispheric specialization for processing auditory non-speech stimuli. *Cereb Cortex,* 16, 1266-1275.

Jancke, L., Wustenberg, T., Scheich, H. & Heinze, H. J. (2002), Phonetic perception and the temporal cortex, *Neuroimage,* 15, 733-746.

Joanisse, M. F. & Gati, J. S. (2003), Overlapping neural regions for processing rapid temporal cues in speech and non-speech signals, *Neuroimage,* 19, 64-79.

Jung-Beeman, M. (2005), Bilateral brain processes for comprehending natural language, *Trends Cogn Sci,* 9, 512-518.

Kaan, E. & Swaab, T. Y. (2002), The brain circuitry of syntactic comprehension, *Trends Cogn Sci,* 6, 350-356.

Knaus, T. A., Bollich, A. M., Corey, D. M., Lemen, L. C. & Foundas, A. L. (2006), Variability in perisylvian brain anatomy in healthy adults, *Brain Lang,* 97, 219-232.

Koechlin, E. & Jubault, T. (2006), Broca's area and the hierarchical organization of human behavior, *Neuron,* 50, 963-974.

Lattner, S., Meyer, M. & Friederici, A. D. (2005), Voice perception: Sex, pitch, and the right hemisphere, *Hum Brain Mapp,* 24, 11-20.

LeMay, M. (1982), Morphological Aspects of Human-Brain Asymmetry – an Evolutionary Perspective, *Trends Neurosci,* 5, 273-275.

Liegeois-Chauvel, C., Giraud, K., Badier, J. M., Marquis, P. & Chauvel, P. (2001), Intracerebral evoked potentials in pitch perception reveal a functional asymmetry of the human auditory cortex, *Ann N Y Acad Sci.,* 930, 117-132.

Liegeois-Chauvel, C., de Graaf, J. B., Laguitton, V. & Chauvel, P. (1999), Specialization of left auditory cortex for speech perception in man depends on temporal coding, *Cereb Cortex,* 9, 484-496.

Lindenberg, R., Fangerau, H. & Seitz, R. J. (2007), 'Broca's area' as a collective term?, *Brain Lang,* 102, 22-29.

Luo, H. & Poeppel, D. (2007), Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex, *Neuron,* 54, 1001-1010.

Marcus, G. F., Vouloumanos, A. & Sag, I. A. (2003), Does Broca's play by the rules?, *Nat Neurosci,* 6, 651-652.

Meyer, M., Alter, K. & Friederici, A. (2003), Functional MR imaging exposes differential brain responses to syntax and prosody during auditory sentence comprehension, *Journal of Neurolinguistics,* 16, 277-300.

Meyer, M., Alter, K., Friederici, A. D., Lohmann, G. & von Cramon, D. Y. (2002), FMRI reveals brain regions mediating slow prosodic modulations in spoken sentences, *Hum Brain Mapp,* 17, 73-88.

Meyer, M., Baumann, S. & Jancke, L. (2006), Electrical brain imaging reveals spatio-temporal dynamics of timbre perception in humans. *Neuroimage,* 32, 1510-1523.

Meyer, M., Baumann, S., Wildgruber, D. & Alter, K. (2007), How the brain laughs. Comparative evidence from behavioral, electrophysiological and neuroimaging studies in human and monkey, *Behav Brain Res,* 182, 245-260.

Meyer, M. & Jancke, L. (2006), Involvement of the left and right frontal operculum in speech and non-speech perception and production, In *Broca's region* (Y. Grodzinsky & K. Amunts, editors), New York: Oxford University Press, 218-241.

Meyer, M., Steinhauer, K., Alter, K., Friederici, A. D. & von Cramon, D. Y. (2004), Brain activity varies with modulation of dynamic pitch variance in sentence melody, *Brain Lang,* 89, 277-289.

Meyer, M., Toepel, U., Keller, J., Nussbaumer, D., Zysset, S. & Friederici, A. D. (2007), Neuroplasticity of sign language: implications from structural and functional brain imaging, *Restor Neurol Neurosci,* 25, 335-351.

Meyer, M., Zaehle, T., Gountouna, V. E., Barron, A., Jancke, L. & Turk, A. (2005), Spectro-temporal processing during speech perception involves left posterior auditory cortex, *Neuroreport,* 16, 1985-1989.

Meyer, M., Zysset, S., von Cramon, D. Y. & Alter, K. (2005), Distinct fMRI responses to laughter, speech, and sounds along the human peri-sylvian cortex, *Brain Res Cogn Brain Res,* 24, 291-306.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T. & Zilles, K. (2001), Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system, *Neuroimage,* 13, 684-701.

Nucifora, P. G., Verma, R., Melhem, E. R., Gur, R. E. & Gur, R. C. (2005), Leftward asymmetry in relative fiber density of the arcuate fasciculus, *Neuroreport,* 16, 791-794.

Ochiai, T., Grimault, S., Scavarda, D., Roch, G., Hori, T., Riviere, D., Mangin, J. F. & Regis, J. (2004), Sulcal pattern and morphology of the superior temporal sulcus, *Neuroimage,* 22, 706-719.

Oechslin, M., Meyer, M., Jäncke, L. (2009), Absolute pitch – functional evidence of speech-relevant auditory acuity, Accepted for Publication in *Cerebral Cortex.*

Okamoto, H., Stracke, H., Draganova, R. & Pantev, C. (2009), Hemispheric Asymmetry of Auditory Evoked Fields Elicited by Spectral versus Temporal Stimulus Change, *Cerebral Cortex,* Epub ahead of Print.

Pell, M. D., Cheang, H. S. & Leonard, C. L. (2006), The impact of Parkinson's disease on vocal-prosodic communication from the perspective of listeners, *Brain Lang,* 97, 123-134.

Petrides, M. (2006), Broca's area in the human and the non-human primate brain, in *Broca's region* (Y. Grodzinsky & K. Amunts, editors), New York: Oxford University Press, 31-46.

Phillips, D. P. & Farmer, M. E. (1990), Acquired word deafness, and the temporal grain of sound representation in the primary auditory cortex, *Behav Brain Res,* 40, 85-94.

Phillips, D. P., Taylor, T. L., Hall, S. E., Carr, M. M. & Mossop, J. E. (1997), Detection of silent intervals between noises activating different perceptual channels: some properties of 'central' auditory gap detection, *J Acoust Soc Am,* 101, 3694-3705.

Poeck, K. & Pietron, H. P. (1981), The influence of stretched speech presentation on token test performance of aphasic and right brain damaged patients, *Neuropsychologia,* 19, 133-136.

Poeppel, D. (2001), Pure word deafness and the bilateral processing of the speech code, *Cognitive Science,* 25, 679-693.

Poeppel, D. (2003), The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time', *Speech Communication,* 41, 245-255.

Poeppel, D. & Embick, D. (2005), Defining the relation between linguistics and neuroscience. In *Twenty-first century psycholinguistics. Four cornerstones* (A. Cutle, editor), Mahwah (NJ): Lawrence Erlbaum, 103-118.

Poeppel, D., Guillemin, A., Thompson, J., Fritz, J., Bavelier, D. & Braun, A. R. (2004), Auditory lexical decision, categorical perception, and FM direction discrimination differentially engage left and right auditory cortex, *Neuropsychologia,* 42, 183-200.

Poeppel, D. & Hickok, G. (2004), Towards a new functional anatomy of language, *Cognition,* 92, 1-12.

Poeppel, D., Idsardi, W. J. & van Wassenhove, V. (2008), Speech perception at the interface of neurobiology and linguistics, *Philos Trans R Soc Lond B Biol Sci,* 363, 1071-1086.

Price, C. J. (2000), The anatomy of language: contributions from functional neuroimaging, *J Anat,* 197, 335-359.

Price, C. J., Gorno-Tempini, M. L., Graham, K. S., Biggio, N., Mechelli, A., Patterson, K. & Noppeney, U. (2003), Normal and pathological reading: converging data from lesion and imaging studies, *Neuroimage,* 20 (Suppl 1), 30-41.

Price, C. J. & Mechelli, A. (2005), Reading and reading disturbance, *Curr Opin Neurobiol,* 15, 231-238.

Pujol, J., Deus, J., Losilla, J. M. & Capdevila, A. (1999), Cerebral lateralization of language in normal left-handed people studied by functional MRI, *Neurology,* 52, 1038-1043.

Riecker, A., Wildgruber, D., Dogil, G., Grodd, W. & Ackermann, H. (2002), Hemispheric lateralization effects of rhythm implementation during syllable repetitions: an fMRI study, *Neuroimage,* 16, 169-176.

Rilling, J. K., Glasser, M. F., Preuss, T. M., Ma, X., Zhao, T., Hu, X., Behrens, T. E. (2008), The evolution of the arcuate fasciculus revealed with comparative DTI, *Nat Neurosci,* 11, 382-384.

Rimol, L. M., Specht, K., Weis, S., Savoy, R. & Hugdahl, K. (2005), Processing of sub-syllabic speech units in the posterior temporal lobe: an fMRI study, *Neuroimage,* 26, 1059-1067.

Saberi, K. & Perrott, D. R. (1999), Cognitive restoration of reversed speech, *Nature,* 398, 760.

Sandmann, P., Eichele, T., Buechler, M., Debener, S., Jäncke, L., Diellier, N., Hugdahl, K., Meyer, M. (2009), Evaluation of evoked potentials to dyadic tones after cochlear implantation, *Brain Epub* ahead of Print.

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M. S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J. & Weiller, C. V. (2008), Ventral and dorsal pathways for language, *Proc Natl Acad Sci USA,* 105, 18035-18040.

Schlaug, G., Jancke, L., Huang, Y. & Steinmetz, H. (1995), In vivo evidence of structural brain asymmetry in musicians, *Science,* 267, 699-701.

Schmahmann, J. D., Pandya, D. N., Wang, R., Dai, G., D'Arceuil, H.E., de Crespigny, A. J., Wedeen, V. J. (2007), Association fibre pathways of the brain: parallel observations from diffusion spectrum imaging and autoradiography, *Brain,* 130, 630-653.

Schmidt, C. F., Zaehle, T., Meyer, M., Geiser, E., Boesiger, P. & Jancke, L. (2008), Silent and continuous fMRI scanning differentially modulate activation in an auditory language comprehension task, *Hum Brain Mapp 29*, 46-56.

Schonwiesner, M., Rubsamen, R. & von Cramon, D. Y. (2005), Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex, *Eur J Neurosci,* 22, 1521-1528

Schwartz, J. & Tallal, P. (1980), Rate of acoustic change may underlie hemispheric specialization for speech perception, *Science,* 207, 1380-1381.

Scott, S. K. & Wise, R. J. (2004), The functional neuroanatomy of prelexical processing in speech perception, *Cognition,* 92, 13-45.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J. & Ekelid, M. (1995), Speech recognition with primarily temporal cues, *Science,* 270, 303-304.

Shattuck-Hufnagel, S. & Turk, A. E. (1996), A prosody tutorial for investigators of auditory sentence processing, *J Psycholinguist Res,* 25, 193-247.

Sidtis, J. J. (2007), Some problems for representations of brain organization based on activation in functional imaging, *Brain Lang,* 102, 130-140.

Sigalovsky, I. S., Fischl, B. & Melcher, J. R. (2006), Mapping an intrinsic MR property of gray matter in auditory cortex of living humans: a possible marker for primary cortex and hemispheric differences, *Neuroimage,* 32, 1524-1537.

Sonntag, G. P. & Portele, T. (1998), PURR – a method for prosody evaluation and investigation, *Computer Speech and Language,* 12, 437-451.

Sowell, E. R., Thompson, P. M., Rex, D., Kornsand, D., Tessner, K. D., Jernigan, T. L. & Toga, A. W. (2002), Mapping sulcal pattern asymmetry and local cortical surface gray matter distribution in vivo: maturation in perisylvian cortices, *Cereb Cortex,* 12, 17-26.

Specht, K., Holtel, C., Zahn, R., Herzog, H., Krause, B. J., Mottaghy, F. M., Radermacher, I., Schmidt, D., Tellmann, L. & Weis, S. (2003), Lexical decision of nonwords and pseudowords in humans: a positronemission tomography study, *Neurosci Lett,* 345, 177-181.

Steinmetz, H. (1996), Structure, functional and cerebral asymmetry: in vivo morphometry of the planum temporale, *Neurosci Biobehav Rev,* 20, 587-591.

Steinschneider, M., Volkov, I. O., Fishman, Y. I., Oya, H., Arezzo, J. C. & Howard, M. A., 3rd. (2005), Intracortical responses in human and monkey primary auditory cortex support a temporal processing mechanism for encoding of the voice onset time phonetic parameter, *Cereb Cortex,* 15, 170-186.

Stowe, L. A., Haverkort, M. & Zwarts, F. (2005), Rethinking the neurological basis of language, *Lingua,* 115, 997-1045.

Su, P., Kuan, C., Kaga, K., Sano, M. & Mima, K. (2008), Myelination progression in language-correlated regions in brain of normal children determined by quantitative MRI assessment, *International Journal of Pediatric Otorhinolaryngology,* 72, 1751-1763.

Tallal, P., Miller, S. & Fitch, R. H. (1993), Neurobiological basis of speech: a case for the preeminence of temporal processing, *Ann N Y Acad Sci,* 682, 27-47.

Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S. S., Schreiner, C., Jenkins, W. M. & Merzenich, M. M. (1996), Language comprehension in language-learning impaired children improved with acoustically modified speech, *Science,* 271, 81-84.

Tettamanti, M. & Weniger, D. (2006), Broca's area: a supramodal hierarchical processor?, *Cortex,* 42, 491-494.

Toga, A. W. & Thompson, P. M. (2003), Mapping brain asymmetry, *Nat Rev Neurosci,* 4, 37-48.

Trebuchon-Da Fonseca, A., Giraud, K., Badier, J. M., Chauvel, P. & Liegeois-Chauvel, C. (2005), Hemispheric lateralization of voice onset time (VOT) comparison between depth and scalp EEG recordings, *Neuroimage,* 27, 1-14.

Uppenkamp, S., Johnsrude, I. S., Norris, D., Marslen-Wilson, W. & Patterson, R. D. (2006), Locating the initial stages of speech-sound processing in human temporal cortex, *Neuroimage,* 31, 1284-1296.

Van Lancker Sidtis, D. (2006), Does functional neuroimaging solve the questions of neurolinguistics?, *Brain Lang,* 98, 276-290.

Van Lancker Sidtis, D., Pachana, N., Cummings, J. L. & Sidtis, J. J. (2006), Dysprosodic speech following basal ganglia insult: toward a conceptual framework for the study of the cerebral representation of prosody, *Brain Lang,* 97, 135-153.

Vernooij, M. W., Smits, M., Wielopolski, P. A., Houston, G. C., Krestin, G. P. & van der Lugt, A. (2007), Fiber density asymmetry of the arcuate fasciculus in relation to functional hemispheric language lateralization in both right- and left-handed healthy subjects: a combined fMRI and DTI study, *Neuroimage,* 35, 1064-1076.

Vigneau, M., Beaucousin, V., Herve, P. Y., Duffau, H., Crivello, F., Houde, O., Mazoyer, B. & Tzourio-Mazoyer, N. (2006), Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing, *Neuroimage,* 30, 1414-1432.

Von Steinbuchel, N. (1998), Temporal ranges of central nervous processing: clinical evidence, *Exp Brain Res,* 123, 220-233.

Warren, J. D., Scott, S. K., Price, C. J. & Griffiths, T. D. (2006), Human brain mechanisms for the early analysis of voices, *Neuroimage,* 31, 1389-1397.

Wernicke, C. (1874), *Der aphasische Symptomkomplex: eine psychologische Studie auf anatomischer Basis*, Breslau: Cohn & Weigert.

Willems, R. M. & Hagoort, P. (2007), Neural evidence for the interplay between language, gesture, and action: a review, *Brain Lang,* 101, 278-289.

Wise, R. J., Scott, S. K., Blank, S. C., Mummery, C. J., Murphy, K. & Warburton, E. A. (2001), Separate neural subsystems within 'Wernicke's area', *Brain,* 124, 83-95.

Zaehle, T., Geiser, E., Alter, K., Jancke, L. & Meyer, M. (2008), Segmental processing in the human auditory dorsal stream, *Brain Res*, 1220, 179-190.

Zaehle, T., Jancke, L., Hermann, C. S. & Meyer, M. (2009), Pre-attentive Spectro-temporal Feature Processing in the Human Auditory System, *Brain Topography,* Epub ahead of Print.

Zaehle, T., Jancke, L. & Meyer, M. (2007), Electrical brain imaging evidences left auditory cortex involvement in speech and non-speech discrimination based on temporal features. *Behav Brain Funct,* 3, 63.

Zaehle, T., Schmidt, C. F., Meyer, M., Baumann, S., Baltes, C., Boesiger, P. & Jancke, L. (2007), Comparison of 'silent' clustered and sparse temporal fMRI acquisitions in tonal and speech perception tasks, *Neuroimage,* 37, 1195-1204.

Zaehle, T., Wustenberg, T., Meyer, M. & Jancke, L. (2004), Evidence for rapid auditory perception as the foundation of speech processing: a sparse temporal sampling fMRI study, *Eur J Neurosci,* 20, 2447-2456.

Zatorre, R. J. & Belin, P. (2001), Spectral and temporal processing in human auditory cortex, *Cereb Cortex,* 11, 946-953.

Zatorre, R. J., Belin, P. & Penhune, V. B. (2002), Structure and function of auditory cortex: music and speech, *Trends Cogn Sci,* 6, 37-46.

PHONETIC CONTRASTS IN FOREIGN LANGUAGE PERCEPTION: A NEUROPSYCHOLOGICAL STUDY ON SERBIAN AFFRICATES

Nuria Kaufmann [a], Martin Meyer [a], Stephan Schmid [b]
[a] Institute of Psychology (Division of Neuropsychology), University of Zurich
[b] Phonetics Laboratory, University of Zurich
nuria_kaufmann@access.uzh.ch, m.meyer@psychologie.uzh.ch, schmidst@pholab.uzh.ch

1. ABSTRACT

This study addresses the question to which extent phonetic contrasts of a foreign language are perceived more easily by speakers of a native language that shares similar phonetic categories. The focus lies on two postalveolar and two alveolo-palatal affricates of Serbian: [tʃ] (postalveolar, voiceless), [tɕ] (alveolo-palatal, voiceless), [dʒ] (postalveolar, voiced) and [dʑ] (alveolo-palatal, voiced). Swiss-German dialects have the postalveolar voiceless affricate [tʃ] only, while the Rhaeto-Romance variety of *Sursilvan* has three different affricates, i.e. [tʃ], [tɕ], and [dʑ].

In a EEG experiment using a Multi-Deviant Mismatch Negativity (MMN) paradigm, 15 Swiss-German speaking adults and 15 Rhaeto-Romance speaking adults between the ages of 20 to 30 years were instructed to focus on a random reading while not paying attention to the auditory stimuli. The hypothesis is a significant difference in processing between the two groups: Swiss-German speakers will not be able to reliably distinguish the four Serbian affricates. Rhaeto-Romance speakers on the other hand are expected to be able to distinguish all four affricates as they share three of the four phonetic categories.

A significant group-effect was found to corroborate that Rhaeto-Romance speakers process the Serbian affricates differently from the Swiss-German speakers.

2. INTRODUCTION

There is a diversified discussion on how and when we best learn a foreign language (L2). Some advocate that foreign-language learning is no longer possible without any accent after a 'Critical Period' (e.g., Lenneberg, 1967; Kuhl, 2004). Others plead in favor of a continuous mode of foreign-language learning which does not differ significantly between children and adults (e.g. Friederici, 2005). The Critical Period Hypothesis states that an L2 exhibits different processing patterns than the L1. A 'less is more' Hypothesis on the other hand states that processing patterns could be similar, provided the new grammar to be learnt is small (Friederici *et al.*, 2002). This would conform to the assumption that language competence in the L2 affects processing patterns more significantly than age of acquisition (e.g. Winkler *et al.*, 1999).

A number of neuropsychological studies reveal an improved ability to discriminate foreign language sounds with higher language proficiency (e.g. Winkler *et al.*, 1999). Mismatch negativity paradigms have shown that fluent non-native speakers develop a cortical memory for the foreign language phonemes (Näätänen *et al.*, 1997 and Winkler *et al.*, 1999). Such recognition patterns presumably develop gradually with the exposure to the new language. Even in a well-learned second language, however, phoneme representations of the native language were found to exert a strong influence on contrast detection

(Nenonen *et al.*, 2005). Consequently, different mother tongues (L1s) could out-fit one differently to learn a certain foreign language. Thus, we consider the MMN approach most suitable to address our question at issue.

The mismatch negativity (MMN) is a negative deflection that peaks approximately 100-250 ms after the stimulus onset. Classically, the MMN is elicited in the so-called 'oddball paradigm' (see below) as response to sudden changes (deviants) in an auditory sequence (usually represented by standard stimuli). The MMN is understood as a pre-attentive, automatic response. Nevertheless, its amplitude can be enhanced under attention (Näätänen *et al.*, 2004).

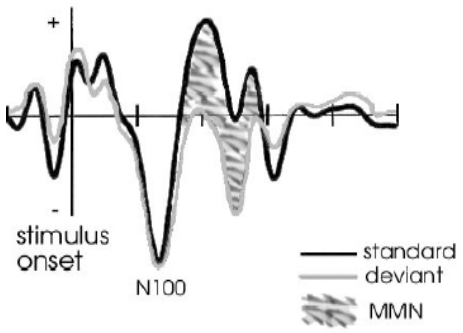


Figure 1: Schematic illustration of the oddball paradigm and the resulting MMN (Lipski, 2006: 45)

The MMN can be observed in the difference wave that is obtained by subtracting the Event-related Potential (ERP) of the standard-stimulus from the deviant-ERP. As the MMN arises only as a response to a new event, the formation of a memory trace to the standard stimuli is preconditioned (Titova & Näätänen, 2001). Incoming deviant stimuli are compared to the regular pattern of the standard sound. Näätänen and colleagues (1997) find enhanced MMN responses to phoneme changes that are relevant to the subject's native language. Reflecting the processing of abstract regularities, long-term memory traces and learning effects, MMN is therefore applicable for the study of cognitive functions.

Up to now, MMN experiments have been conducted on various syllable types (compare Näätänen *et al.*, 1997; Lipski, 2006), but the considerable variety of affricate categories across the languages of the world (Ladefoged & Maddieson, 1996: 90-91) calls for advanced research also on this specific topic. Let us therefore briefly illustrate the affricate subsystems of the three languages involved in the present study, i.e. Serbian, Rhaeto-Romance (Sursilvan), and Swiss-German.

The consonant inventory of Serbian is rather complex (Corbett, 1987: 396). Within the manner of articulation of affricates, Serbian differentiates four categories that are used in our experiment, namely [tʃ] (postalveolar, voiceless), [tɕ] (alveolo-palatal, voiceless), [dʒ] (postalveolar, voiced), and [dʑ] (alveolo-palatal, voiced); it has also has [ts] (alveolar, voiceless) which is not part of the experiment. Phoneticians disagree, however, whether the palatal obstruents are affricates or stops; moreover, they are sometimes described as palatal, sometimes as alveolo-palatal (Morén, 2006).

The Rhaeto-romance language territory in Switzerland is divided into five dialects (Haiman & Benincà, 1992): *Sursilvan, Sutsilvan, Sumiran, Puter* and *Vallader*. For our experiment, *Sursilvan* has been chosen, because it is the most spoken dialect. *Sursilvan* shares three of the four affricates with Serbian, namely [tʃ], [tɕ] and [dʑ] (Liver, 1999). Again, scholars disagree with regard to the phonetic description of [tɕ] and [dʑ], which are classified either as stops or as affricates on the one hand, and as palatal or as palato-alveolar on the other (compare Brunner, 1963; Schmid, 2010). For the purpose of this study, we consider them as alveolo-palatal affricates, just as the Serbian ones. *Sursilvan* [tʃ] also seems to share the typical lip rounding of Serbian (Morén, 2006), but it lacks the voiced postalveolar affricate [dʒ] that exists in Serbian (as well as in Italian and in the Rhaeto-Romance dialects of the Engadine).

Now turning to Swiss-German, we might illustrate its consonant inventory by referring to the Zurich dialect, which contains four voiceless affricates, namely labial [pf], alveolar [ts], postalveolar [tʃ], and velar (sometimes uvular) [kx] (Fleischer & Schmid, 2006). Thus, out of the manner of articulation we are interested in, Swiss-German only has [tʃ]. Considering that the four affricates in our study differ in voicing and place of articulation, it must be pointed out that Swiss-German speakers do not differentiate contrasts of the affricate category in either dimension. However, due to some knowledge of English (which has both a voiceless and a voiced postalveolar affricate), the distinction of voicing might be easier for them than the detection of another place of articulation.

Affricates	post-alveolar (voiceless)	alveolo-palatal (voiceless)	alveolo-palatal (voiced)	postalveolar (voiced)
Serbian	[tʃ]	[tɕ]	[dʑ]	[dʒ]
Rhaeto-Romance	[tʃ]	[tɕ]	[dʑ]	–
Swiss-German	[tʃ]	–	–	–

Table 1: Affricates in Serbian, Rhaeto-Romance and Swiss-German (comparison of the phoneme categories that are relevant in this study)

3. MATERIALS AND METHODS

3.1 Stimuli

The four Serbian syllables [tɕa], [dʑa], [dʒa], and [tʃa] served as stimuli in the Electro-Encephalogram (EEG) recording (see 3.3). The usage of CV (consonant-vowel) syllables was motivated by the fact that isolated affricates, especially voiceless ones resemble nonspeech noise. This impression is reinforced by the repetitive presentations that are necessary in EEG experiments. Furthermore, the transitions to subsequent vowels may provide important perceptual cues for the identification of the affricate. Because the vowel [a] is universally unmarked, we decided to apply this vowel. In contrast to [u] and [o], [a] does not lead to anticipatory lip rounding during the production of the affricate and there is no coarticulatory influence of a palatal glide for [a].

The stimuli used for the experiment were digitally recorded in a sound proof chamber at the Phonetics Laboratory of the University of Zurich. A sampling rate of 44100 Hz and 16 bit quantization were used. A female native speaker of Serbian read the four syllables aloud in twelve variations each: they were spoken three times in a CV sequence, in a VCV sequence and in an existing Serbian word (*ćaskati* "to chat", *čarapa* "sock", *đavol* "devil", *džaba* "frog"). All of the syllables that served as acoustic stimuli were pronounced inside a carrier phrase where the preceding segment was a vowel (*Prvo ća, drugo ća, treće ća* "first ća, second ća, third ća"), which allowed us to precisely detect the starting point of the consonant under examination.

3.1.1 Acoustic analysis

In a first step, the duration of the closure and the release phase of 48 affricates was measured manually on the basis of an introspection of the wave forms and spectrograms provided by *Praat* (Boersma & Weenink, 2009); after that, the duration of the whole syllable was noted. In order to guarantee a certain reliability of the measurements, the procedure was repeated in order to obtain two times 24 tokens, including four stimuli – each in three repetitions (see carrier phrase) and two conditions (CV and VCV).

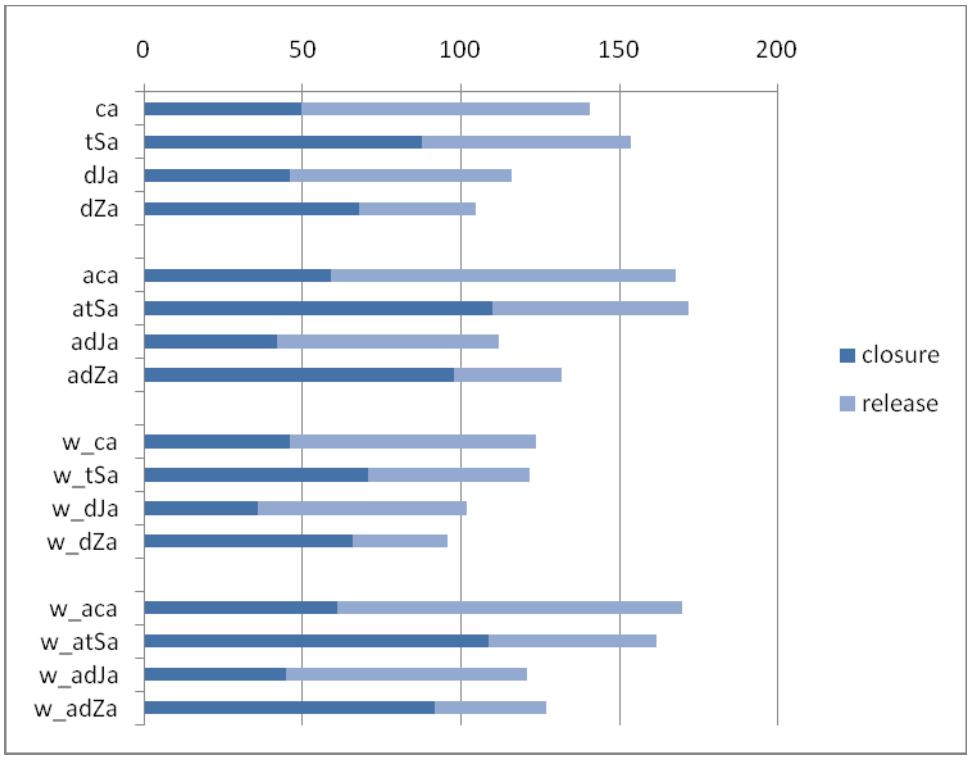


Figure 2: Mean duration values (ms) for the closure and the release phase of the affricates in logatomes (up) and Serbian words (down)

Figure 2 illustrates the mean duration values (ms) for the closure and the release phase of of the affricates in the recorded career sentences. The upper part of the graph shows the mean duration values of the affricates [tɕ], [tʃ], [dʒ] and [dʑ] pronounced in CV and CVC logatoms;[1] the lower part shows the mean duration of the same affricates pronounced in Serbian words.

In all four contexts, voicing clearly affects duration, since the two voiceless affricates are always longer than the voiced ones. As regards place of articulation, it results that the two alveolo-palatal affricates always display a relatively shorter closure phase and a longer release phase than the two postalveolar affricates.

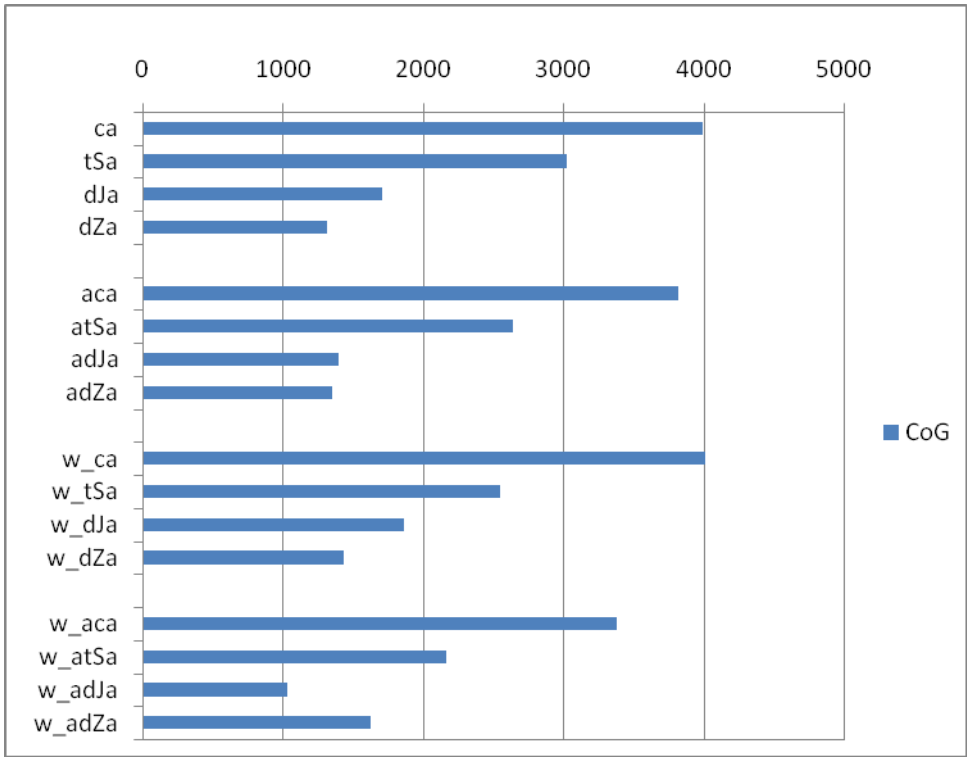


Figure 3: Mean values of the Centre of Gravity (CoG) of each affricate (Hz)

In order to obtain a measure for the spectral characteristics of the affricates, the 'Centre of Gravity' (CoG) was calculated (Forrest *et al.*, 1988; Gordon *et al.*, 2002), using the apposite function in *Praat* (compare Mele & Schmid, 2009: 365-367). Figure 3 illustrates the mean CoG values for the affricates [tɕ], [tʃ], [dʒ] and [dʑ] according to the four different phonosyntactic contexts. The results indicate a significant effect of voicing on the Centre of

[1] For technical reasons, here [tɕ] and [dʑ] are referred to by the symbols [c] and [dJ]; similarly, [tS] stands for [tʃ] and [dZ] for [dʒ].

Gravity, given the clearly higher values for the voiceless affricates; obviously, this result reflects the additional presence of energy in the lower frequency range which appears in the spectrum of voiced obstruents. As regards place of articulation, we note a higher CoG for the palatal affricates as opposed to the postalveolar ones. For the time being, we limit ourselves to observe this as an acoustic fact (with a possible auditive effect in the acute-grave dimension), without speculating on the articulatory nature of the sounds involved; possibly, lip rounding is at stake here.

3.1.2 Selection and editing

The four stimuli used in the experiment were selected from the second recording according to the following criteria: Duration for affricate and vowel about 150 ms, even, constant fundamental frequency (F0) trend. Editing included stylizing the pitch using *Praat* 5045 (Boersma & Weenink, 2009) and setting the overall intensity to 70 dB. Normalization was done using the software *Audition.*[2] This did not change the intensity relation between affricates and vowels in the individual syllables. A Butterworth filter was applied as low-pass filter (5000 Hz) to cut background- and click-sounds using *Audition.* At the onset and at the end of the syllables a smooth rising/falling ramp with duration of 10 ms was added (Gaussian filter). F0 was set to a constant value throughout the vowel with respect to initial F0 value. Duration was normalized by clipping the affricate onset and vowel offset so that each syllable had duration of between 120-185 ms. Finally, the vowel of the syllable [tʃa] was stabilized at a length of 92 ms and was used for all four syllables.

The last step was done in full awareness of the loss of information that is provided by the specific transition of the affricate to the following vowel. As described in Recasens & Espinosa (2007: 149), "the duration of the vowel preceding the affricate ought to be strongly related to the duration of the entire affricate and of its closure period. Spanish, English and Italian data reveal indeed that vowel duration compensates for affricate and closure duration but less clearly so or not at all for frication duration, i.e. the vowel shortens as the affricate and its closure period lengthen and vice versa".

Indeed we observed the same effect. In addition to the described dependency of stimulus-length, the formant constellation in the transition from the affricate to the vowel varies between the four affricates. We performed a behavioural pre-experiment with Swiss-German speakers which showed that the isolated syllables are much too easy to distinguish with this information included. Subjects reported they would be able to easily differentiate the stimuli paying attention only to the 'higher' and 'lower' sounding vowels. However, we were interested in their ability to perceive the spectral part of the affricate only. This confirmed the necessity of taking away this stimulus-specific attributes, although it retrenches the naturalness of our stimuli.

After the final editing, three Serbian and three Rhaeto-Romance speakers were asked to judge the syllables for their 'naturalness' and their discriminability (e.g. Nenonen *et al.*, 2005). Serbian speakers could reliably ascribe each syllable; Rhaeto-Romance speakers encountered increased difficulties, yet they clearly made out "three or more" different syllables.

[2] http://www.adobe.com/products/audition/

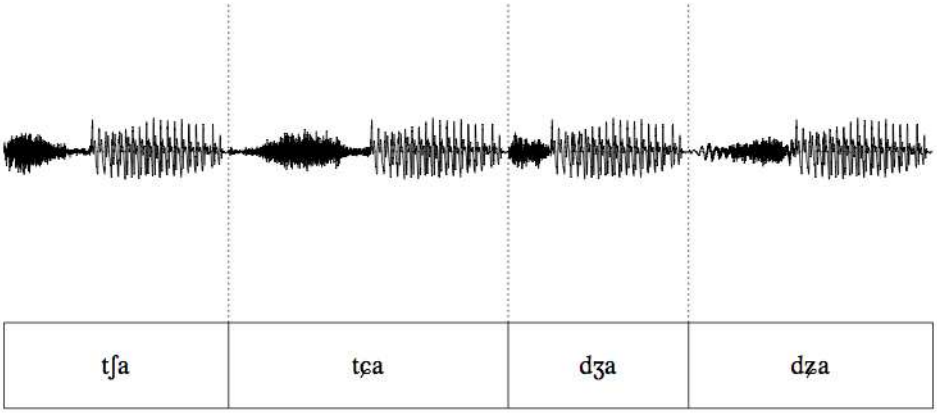


Figure 4: The four Serbian affricates used for the CV stimuli

3.2 Subjects

For the experiment, 30 subjects were recruited: 15 Surselvan mother tongue speakers for the Rhaeto-Romance group and 15 Swiss-German natives. Only righthanded subjects between 20-30 years were assessed. During the installation of the electrodes participants were asked to fill in a questionnaire to file their details: first and foremost their language background (bilingual, second language abilities, etc.). All subjects of both groups learnt English and French in school. Some had knowledge of Spanish. Only four subjects knew Italian. Other second languages were Norwegian (1), Swedish (1), Arabic (2), Hebrew (1) and Latin (3). Knowledge of various second languages might also promote orthographic knowledge which is assumed to influence speech perception (Lipski, 2006). Their contact details were noted in case of further questions.

3.3 Procedure

During the EEG experiment, subjects were seated in an electrically shielded and acoustically attenuated chamber. The data were recorded using a Biosemi active-two amplifier system. 64 active electrodes were installed according to the 10/20 electrode system (Jaspers, 1958) (see figure 5 below).[3] The sampling rate was 512 Hz and impedance was kept below 40 kΛ;[4] vertical and horizontal eye movements were recorded by two bipolar channel pairs placed above and below the left eye, and on the outer canthi of both eyes. For off-line re-referencing, an electrode was attached to the tip of the nose. For head and body movements, participants were monitored through a close-circuit camera system. The whole experiment, including welcoming and hair washing lasted two hours. Subjects received 20 Swiss francs for their participation.

[3] http://www.biosemi.com/headcap.htm
[4] http://www.biosemi.com/faq/shielding%20vs%20active%20electrodes.htm

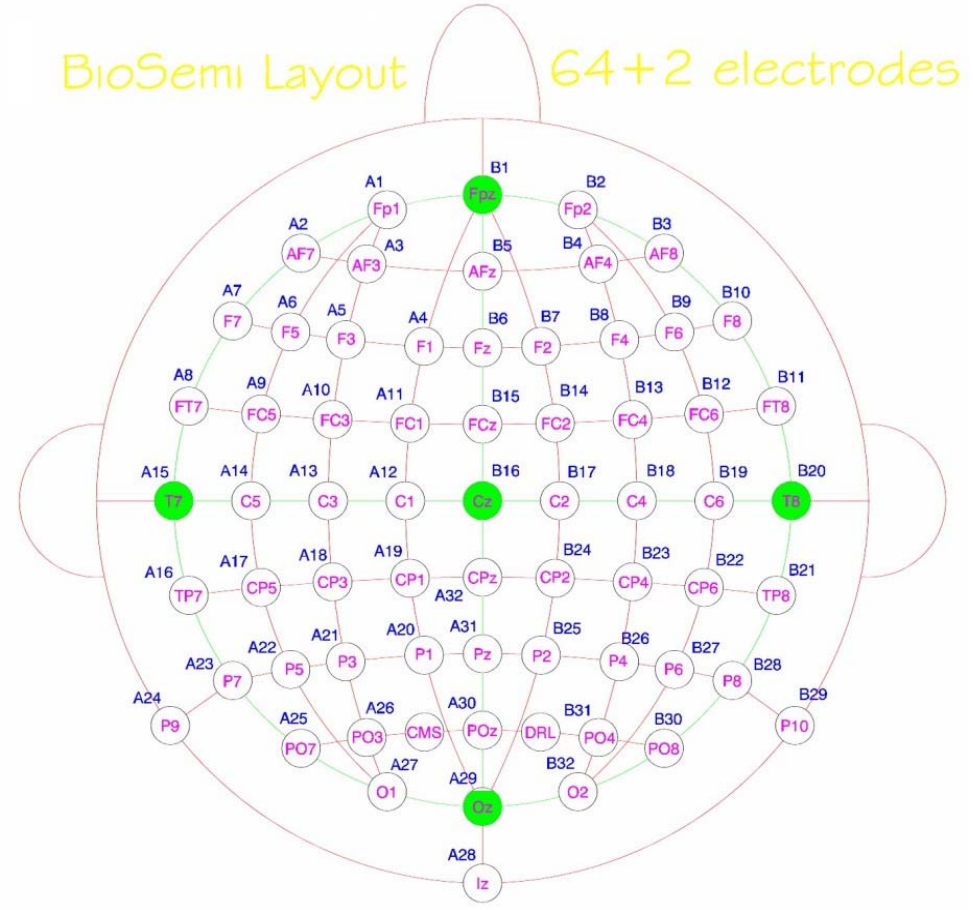


Figure 5: 64 channels 10/20 – layout

3.4 Oddball paradigm

The paradigm follows the idea of Näätänen *et al.* (2004)'s Optimal 1 paradigm. Three 'deviants' (randomly alternating stimuli) are presented alongside the 'standard' (a stimulus which is repeated continuously every second) and not compared individually against the standard as in the classic oddball paradigm. We used an oddball paradigm with 50 percent standard (e.g. [tʃa]) and 50 percent deviant ([tɕa], [dʑa] and [dʒa]) proportion. Furthermore, we used a Multiple-Deviant Paradigm which means that every deviant once acts as the standard. The Inter-stimulus Interval (ISI) was set to 750 ms and the Stimulus Onset Asynchrony (SOA) of 400 ms was jittered. The first two minutes were recorded for closed and open eye-movements (resting EEG). Thereafter, eight passive listening blocks followed. Block sequences were randomized between subjects. Participants were asked to read a random text and not pay attention to the syllables they heard through the head-

phones, but to treat them as 'background music'. At the beginning of each block there were 15 repetitions of the standard to attune the subjects' ear to the respective standard. Therefore, a stimulus block included 150 deviants and 165 standards; in total, 1200 deviant repetitions and 1320 standard repetitions were used, whereof each of the four stimuli appeared 330 times as a standard and 300 times as a deviant. Between blocks, subjects could recess for as long as they wished. On average, breaks lasted three minutes.

3.5 Data Analysis

The data were analyzed using *BrainVision Analyzer 1.05.0000* and *eegLab 6.01* (Matlab). EEGs were offline treated with a 24 dB zero-phase bandpass-filter from 0.1 to 30 Hz. Channels that displayed changes exceeding 150 ⌈V were discarded for further analysis. Unfortunately, we could not use the nose as reference, as the coordinates of this electrode are unknown to the *eegLab system.* Common average Reference (CAR) was therefore applied.

Eye blinks and horizontal movements were corrected by means of independent component analysis (ICA). Due to technical problems while recording, six subjects (three Rhaeto-Romance and three Swiss-German) had to be discarded. EEG recordings were segmented into 600 ms epochs (100 ms pre- and 500 ms post-stimulus) and averaged for each stimulus type separately with 100 ms pre-stimulus as a baseline. ERPs for all stimuli (each stimulus type as a standard and as a deviant) were averaged for each subject and grand-averaged across subjects.

MMN difference waves were computed by subtracting ERPs to the standard from ERPs to the deviant of a chosen stimulus and grand-averaged. Being able to directly compare the response to a certain stimulus acting both as a standard and as a deviant is one of the main advantages of the Multiple-Deviant Paradigm (compare also Grimm *et al.*, 2008). Peak-detection was carried out over a time-window of 180 ms (120-300 ms after stimulus onset).

The presence of the MMN was statistically verified using analysis of variance, one-sampled and independent *t*-tests with *SPSS* at a significance level of 0.05. Analysis involved comparison of groups, stimuli, peaks and latencies. To verify the existence of a true MMN component, activations at Fz were compared with supra-temporal electrodes (TP9 and TP10).[5]

4. RESULTS

Deviant-related MMN potentials were measured by subtracting ERPs elicited by the stimulus operating as a standard sound from ERPs elicited by the same stimulus operating as a deviant sound. This allowed a direct comparison of the physically identical stimulus differing only in its probability of occurrence. ERPs showed orderly N1 and P2 components at central Cz electrode, comparing the two groups (see below Figure 6) and comparing the four stimuli acting as standards for both groups (compare Figure 7 below). Normal distribution was assured with a Kolmogorov Smirnov test.

[5] For position of electrodes see Figure 5.

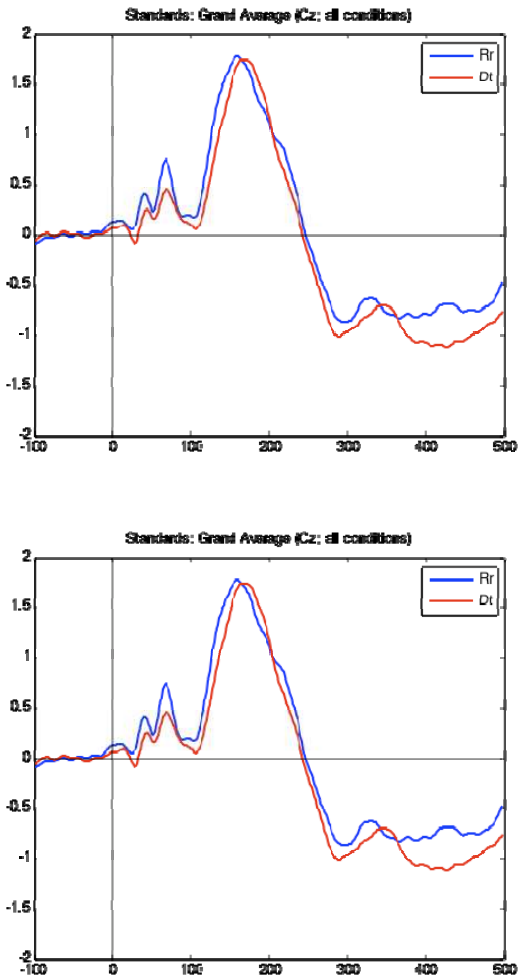


Figure 6: Central Acoustic Event-related Potential (CAERP) plotted at Cz electrode for Rhaeto-Romance speakers (Rr - blue) vs. Swiss-German speakers (Dt - red)

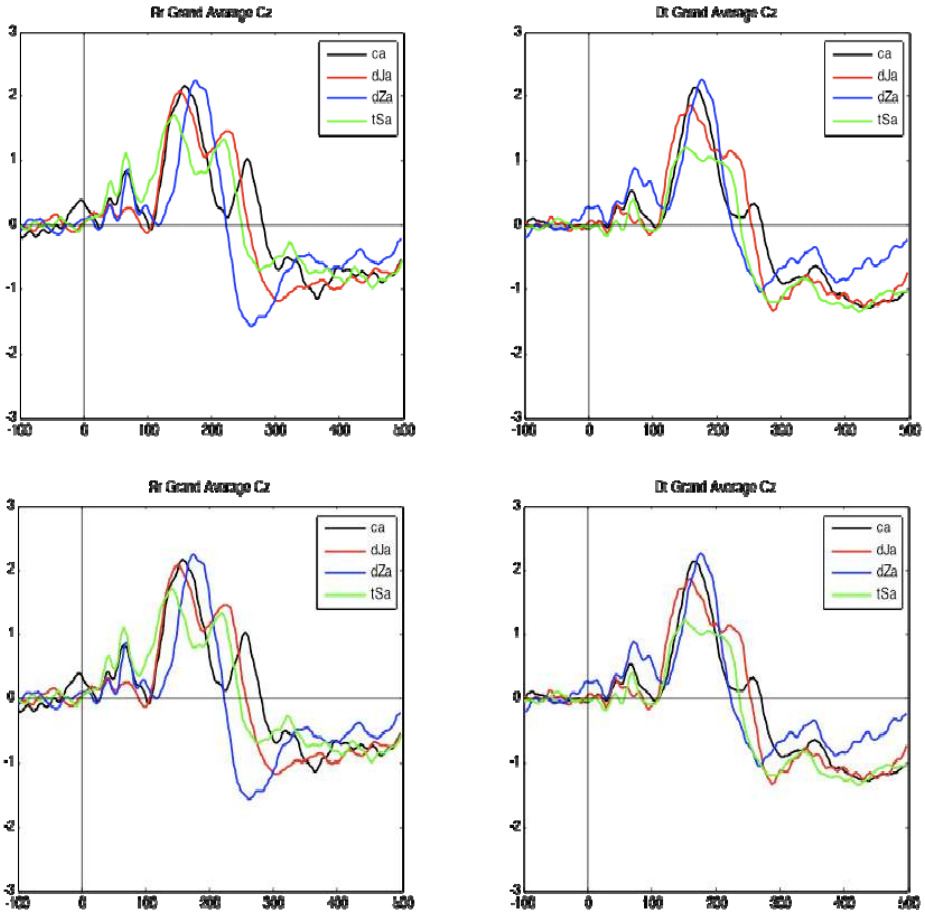


Figure 7: CAERP plotted at Cz electrode
for Rhaeto-Romance speakers (left) vs. Swiss-German speakers (right)[6]

A repeated-measures ANOVA was performed for peaks and latencies separately. The ANOVA included the between-subject factor 'Group' (Rhaeto-Romance vs. Swiss-German), 'Stimulus' ([tʃa], [tɕa], [dʑa] and [dʒa]) and the within-subject factors 'Peak' or 'Latency' (three peak or latency values per stimulus, representing the three deviant conditions). For both language groups in all deviant conditions, negative peaks were observed in the deviant-minus-standard difference waves. The comparison 'Group' x 'Stimulus' x 'Peak' revealed a main effect 'Group' ($p = 0.03$) (compare Table 2 below). As expected, the comparison 'Group' x 'Stimulus' x 'Latency' revealed no main effect.

[6] In the graph, [tɕ] and [dʑ] are referred to by the symbols [c] and [dJ]; similarly, [dZ] stands for [dʒ] and [tS] for [tʃ] ([tɕ] = black, [dʑ] = red, [dʒ] = blue and [tʃ] = green).

		Tests of Between-Subjects Effects (ANOVA)	
		df	Sig. (2-tailed)
Interaction Group:		1	0.027*
Stimulus*Peak*Group:	(Greenhouse-Geisser)	5	0.010***

Table 2: ANOVA: comparing 'Group', 'Stimulus' and 'peaks'

Figure 8 provides the mean scores of each group for all stimuli in the respective deviant conditions; standard deviations are marked with bars and amplitude peaks are compared for each stimulus acting in different deviant positions as compared to acting as a standard. Significant differences that showed in the independent samples t-test are marked with asterisks ($p < 0.001$ = ***, $p < 0.01$ =**, $p < 0.05$ = *); no significant differences were found for the comparison of latency means (see figure 9 below).

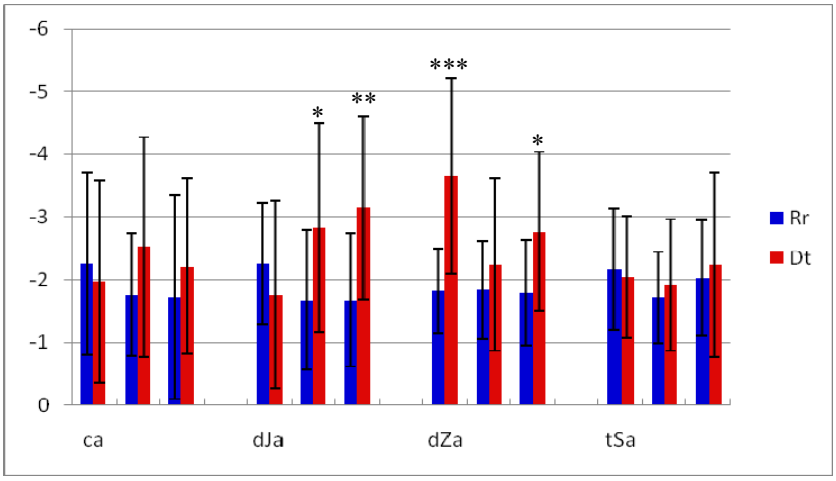


Figure 8: Comparison MMN peak values (µV) and standard deviations
in respective deviant conditions –
Rhaeto-Romance speakers (Rr: blue) and Swiss-German speakers (Dt: red)[7]
(*** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$)

[7] In the graph, [tɕ] and [dʑ] are referred to by the symbols [c] and [dJ]; similarly, [dZ] stands for [dʒ] and [tS] for [tʃ]. The ascribed stimulus represents the standard that was compared to the same stimulus when acting as a deviant.

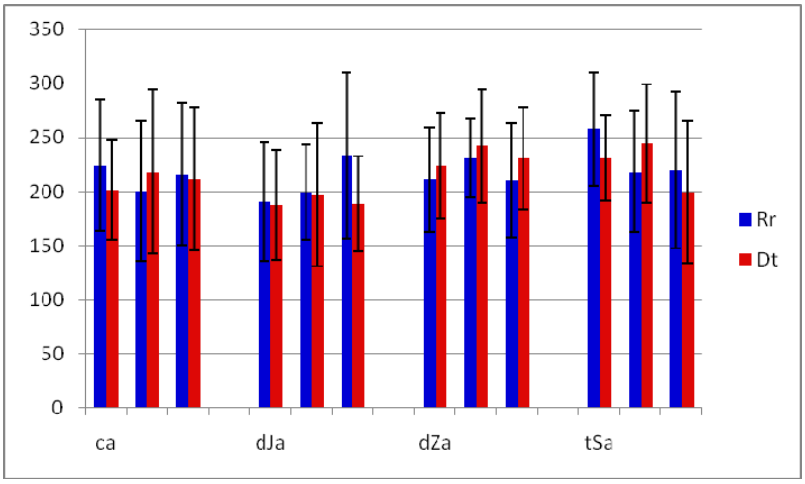


Figure 9: Comparison of groups of MMN latency values (ms) and standard deviations in respective deviant conditions

As expected, the comparison Group * Stimulus * Latency revealed no main effect. Due to slightly different stimulus length (up to 65 ms difference), a systematic latency effect was anticipated. As expected, Tests of Within-Subjects Effects indicated a significant effect for Stimulus (Greenhouse-Geisser $p = 0.013$; not shown in the table), but no interaction with Group.

One-sampled *t*-Tests in both groups for the comparisons of both peaks and latencies were all significant on the $p < 0.001$ level (not shown in the table). Independent Samples *t*-Tests show, that Rhaeto-Romance speakers process phonetic contrasts significantly differently. Surprisingly, the stimulus [tɕa] elicited no significant group difference. Stimulus [dʑa] was processed significantly differently if it served as a deviant beside the standard [tʃa] ($p = 0.01$) and [ʤa] ($p = 0.03$) compared to acting as a standard. Stimulus [ʤa] also displayed significant differences between Rhaeto-Romance and Swiss-German speakers when serving as a deviant in standard blocks [ca] and [tSa] (compare Figure 7 above). Interestingly, these four significant MMN amplitudes are higher for Swiss-German speakers than for Rhaeto-Romance speakers. Based on previous studies with native and non-native phonological contrasts (e.g. Winkler *et al.*, 1999; Peltola *et al.*, 2003; Näätänen *et al.*, 1997), an enhanced MMN for the native-like stimuli was anticipated for the Rhaeto-Romance group. However, in nine out of the twelve contrasts, Swiss-German speakers attained bigger amplitudes than Rhaeto-Romance speakers, four of which were significant (compare Figure 7 above). Surprisingly, no significant group differences were found for the stimulus [tɕa]. As expected, no differences in processing were found for the stimulus [tʃa]. This stimulus is common to speakers of both language groups and should therefore not provoke a significant difference in processing.

The Figures 10 and 11 show the Difference waves Deviant-Standard at Fz electrode for Rhaeto-romance subjects and Swiss-German subjects, respectively. Each stimulus is presented as a standard (blue line) alongside the three different deviants (red lines – left to right) and the resulting difference waves (black lines). The standards from the 1st to the 4th row: [tɕa], [dʑa], [ʤa] and [tʃa].

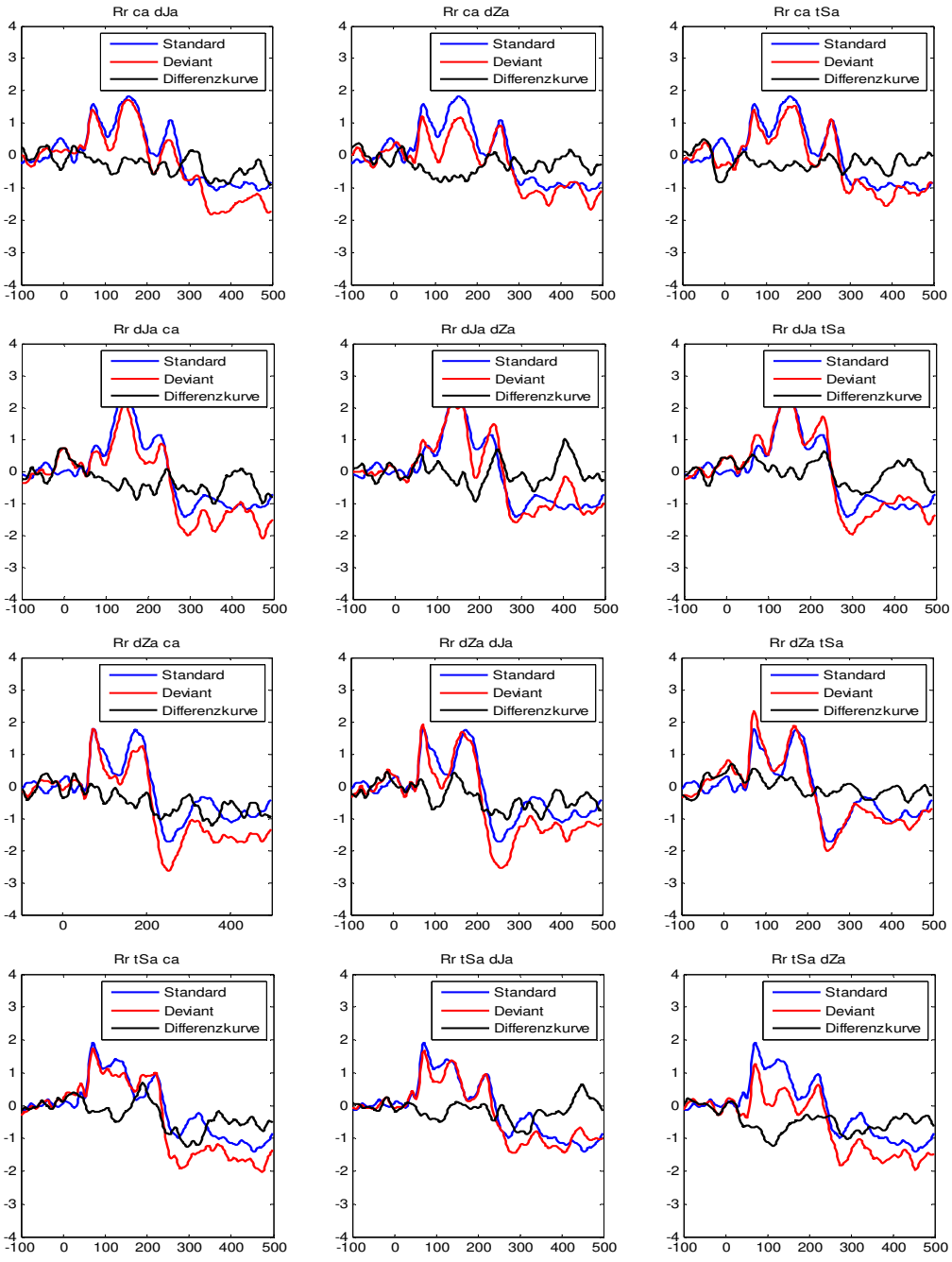


Figure 10: Difference waves Deviant-Standard at Fz electrode for Rhaeto-romance subjects

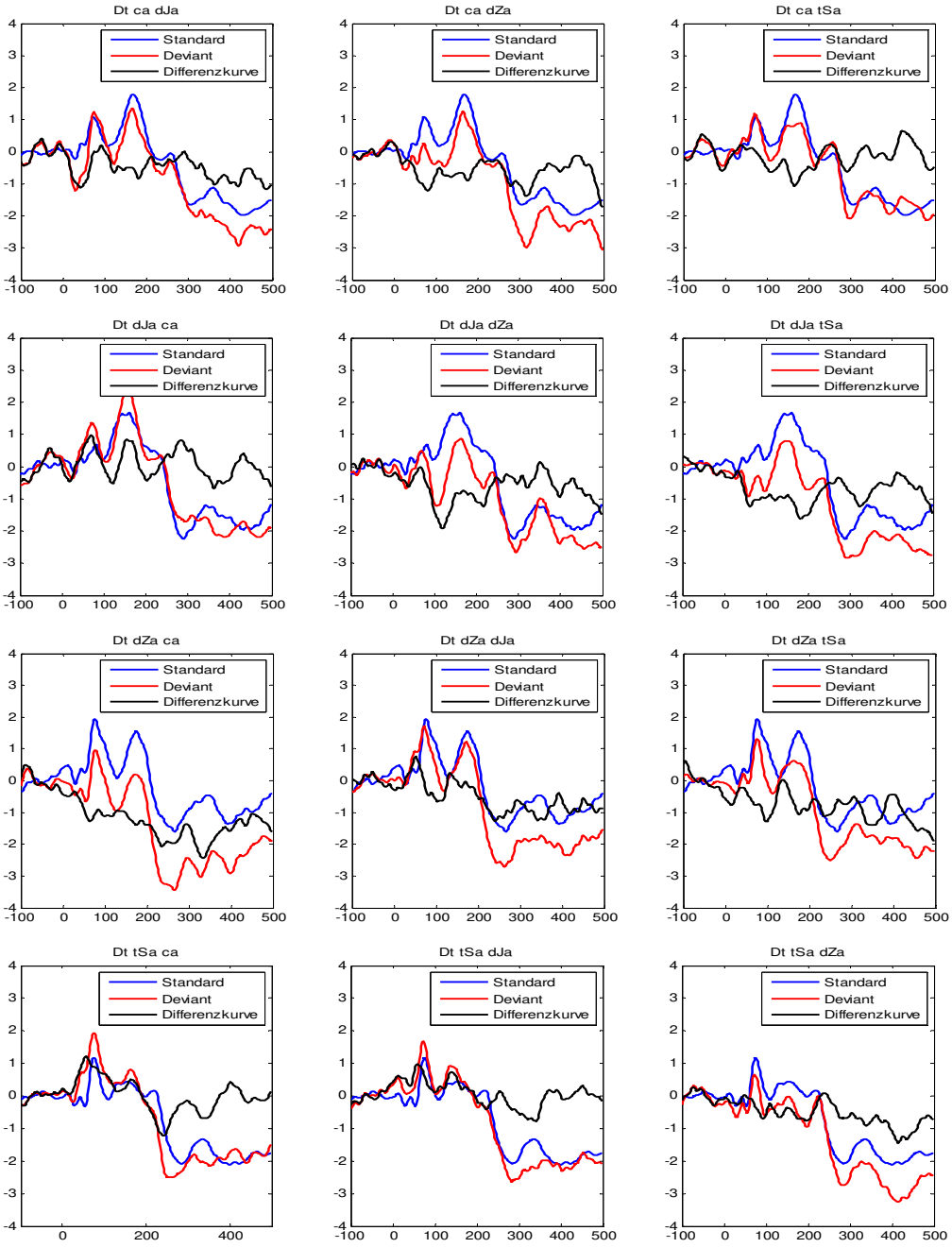


Figure 11: Difference waves Deviant-Standard at Fz electrode for Swiss-German subjects

All MMN curves displayed a typical fronto-central maximum (Fz) with a polarity inversion at the mastoid leads (TP8 & TP7) and latencies between 180-260 ms. The four stimuli are compared in three deviant conditions. For the difference waves, each stimulus is compared in its function as standard to its respective function as a deviant. For example, in the first row of figure 10 (related to the Rhaeto-Romance subjects) and figure 11 (related to the Swiss-German subjects), difference waves of the standard [tɕa] are computed for [tɕa] functioning as a deviant in the first, second and third condition (left to right).

5. DISCUSSION AND CONCLUSION

The overall goals of the MMN experiment were two-fold: first and foremost, to examine the implications of the different language-backgrounds of the two groups, and second to test whether place of articulation or voicing had a stronger influence on the perception of a foreign language phonetic contrast. A significant difference ($p = 0.03$) in MMN amplitudes between groups confirmed a varying way of processing. Half of the expected phonetic contrasts yielded a significant difference between groups. The direction of the effect, however, is unexpected.

With respect to the second goal, we observe a number of interesting findings. What is more, these results challenge the traditional interpretation of the results to the first goal. Instead of the Rhaeto-Romance group, the Swiss-German group shows higher amplitudes. Higher amplitudes in discriminating phonetic contrasts were previously associated with native-like or more proficient processing (compare e.g. Winkler *et al.*, 1999; Näätänen *et al.*,1997). As Rhaeto-Romance shares similar phonetic categories with Serbian, the speakers of this group were expected to out-perform the Swiss-German speakers. We expected higher discrimination ability that would yield larger amplitudes for the Rhaeto-Romance group.

It was hypothesized, that Rhaeto-Romance speakers would be able to differentiate all three stimuli [tɕa], [dʑa] and [dʒa] unknown to Swiss-German speakers. Amplitude differences were expected for contrasts that involved known deviants compared with contrasts that employed an unknown syllable: the Rhaeto-Romance group was expected to show high discrimination amplitudes for deviants [tɕa] and [dʑa] and a lower amplitude for deviant [dʒa]. Except for the deviant [tʃa], lower discrimination amplitudes were expected for the Swiss-German group. Surprisingly, Swiss-German speakers showed higher amplitudes on four out of twelve contrasts that reached significance, even though the contrasts involved deviants that were unknown to them.

Stimulus [dʑa] is known to Rhaeto-Romance speakers. Phoneticians might disagree on the correct phonetic categorization of the Rhaeto-Romance equivalence of the phoneme; nevertheless, the palatal place of articulation as well as high degree of voicing are represented in the phoneme category of interest (not so in Swiss-German, where neither feature is present).

Stimulus [dʒa], on the other hand, is unknown to both Rhaeto-Romance and Swiss-German speakers. Again, both features are known to Rhaeto-Romance speakers. Swiss-German speakers are familiar with the postalveolar affricates, but not with the voiced ones. Possibly, Swiss-German speakers could distinguish the voiced affricates due to previous encounters with English, French and/or Italian.

Stimulus [tʃa] was not expected to reach significance in processing between the two groups as it forms part of the Rhaeto-Romance and the Swiss-German phonetic system alike.

Stimulus [t͡ɕa] might not have yielded a significant result for either group because of too much stimulus editing. It might have been equally difficult for Rhaeto-Romance speakers even though Rhaeto-Romance speakers were able to distinguish the contrast behaviourally which could be interpreted as enhanced performance under the influence of attention.

An initial behavioural rating showed that differences to native sounds are perceivable. The measurement of the pre-attentive and automatic mismatch response confirmed that language background significantly influences the early perception of foreign speech sounds. The direction of the effect, however, was unexpected: Swiss-German speakers displayed higher amplitudes than Rhaeto-Romance speakers on contrasts that are not represented in their native phoneme inventory. There are two possible explanations for this finding. On the hand, a rich second language background in both groups could have evoked memory traces in the Swiss-German speakers as well. Cortical representations of the foreign sound category might have enabled Swiss-German speakers to perform in a comparable if not better way to the Rhaeto-Romance speakers. This would confirm the belief that linguistic experience affects the neural processing window for speech (compare Gandour *et al.*, 2007). On the other hand, overlearning could yield smaller amplitudes to the phonetic contrast in the Rhaeto-Romance group; the higher amplitudes in the Swiss-German subjects would be interpreted as increased neural activity (compare Tervaniemi *et al.*, 2000).

The comparison of two different language groups in their perception of yet another language was relatively unusual. In other studies, either a naive or an advanced group of foreign language learners is compared to native speakers of the language under investigation in their MMN response to phonetic contrasts (e.g. Winkler *et al.*, 1999). Testing the perception abilities of a third language makes the direct comparison of the neural responses difficult as neither of them is a native response.

The naturally spoken Serbian stimuli were stripped of surrounding acoustic cues. This could have played a significant role on the discrimination ability of the Rhaeto-Romance subjects. As Lipski (2006) points out, speakers of languages with an inventory of various fricatives and/or affricates seem to rely more highly on formant transitions than on the frication noise to discriminate these phonemes. Affricate contrasts might need to be placed in their typical context of acoustic cues to be reliably distinguished by native speakers.

Language learning needs to take place in a relevant context – acoustic discrimination ability and knowledge of the related meaning cannot be separated. Phonetic knowledge of the foreign language contrast alone does not enable better discrimination ability if relevant acoustic cues are missing. Thus, it remains unclear whether in our experiment the non-native sounds were assimilated to the native phoneme category or not. Nevertheless, the results support the notion that phonetic features that seem irrelevant to the acquired L1-specific representations are not completely neglected or filtered out. This strongly speaks in favour of the continuous ability to learn foreign language phonemes that are similar/ dissimilar to the L1 phonetic category in adulthood.

6. REFERENCES

Boersma, P. & Weenink, D. (2009), *Praat: doing phonetics by computer*, www.praat.org.

Brunner, R. (1963), Zur Physiologie der Rätoromanischen Affrikaten *tsch* und *tg (ch)*, in *Sprachleben der Schweiz. Sprachwissenschaft, Namenforschung, Volkskunde* (P. Zinsli *et al.*, editors), Bern: Francke, 167-173.

Corbett, G. (1987), Serbo-Croat, in *The world's major languages* (B. Comrie, editor), London: Routledge, 391-409.

Fleischer, J. & Schmid, S. (2006), Zurich German, *Journal of the International Phonetic Association*, 36, 243-253.

Forrest, K., Weismer, G., Milenkovic, P. & R. Dougall (1988), Statistical analysis of word-initial word-final voiceless obstruents: Preliminary data, *Journal of the Acoustical Society of America*, 84, 115-123.

Friederici, A.D., Steinhauer, K. & Pfeifer E. (2002), Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis, *Pnas,* 99, 529-534.

Friederici, A.D. (2005), Neurophysiological markers of early language acquisition: from syllables to sentences, *Trends in Cognitive Sciences,* 9, 481-485.

Gandour, J., Tong, Y., Talavage, T., Wong, D., Dzemidzic, M., Xu, Y. & Lowe, M. (2007), Neural Basis of First and Second Language Processing of Sentence-Level Linguistic Prosody, *Human Brain Mapping,* 28, 94-108.

Gordon, M., Barthmaier, P. & Sands, K. (2002), A cross-linguistic acoustic study of voiceless fricatives*, Journal of the International Phonetic Association,* 32, 141-174.

Grimm, S., Schröger, E., Bendixen, A., Bäss, P., Roye, A. & Deouell, L.Y. (2008), Optimizing the auditory distraction paradigm: Behavioral and event-related potential effects in a lateralized multi-deviant approach, *Clinical Neurophysiology,* 119, 934-947.

Haiman, J. & Benincà, P. (1992), *The Rhaeto-Romance Language,* London: Routledge.

Jaspers, K. (1958), [The physician in the technical age]. *Klin.Wochenschrift,* 36, 1037-1043.

Kuhl, P. (2004), Early Language Acquisition: Cracking the Speech Code, *Nature Reviews,* 5, 831-843.

Ladefoged, P. & Maddieson, I. (1996), *The Sound of the World's Languages*, Oxford: Blackwell.

Lenneberg, E. (1967), *Biological foundations of language*, New York: Wiley.

Lipski, S. C. (2006), *Neural correlates of fricative contrasts across language boundaries,* Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart.

Liver, R. (1999), *Rätoromanisch. Eine Einführung in das Bündnerromanisch,* Tübingen: Narr.

Mele, B. & Schmid, S. (2009), Le occlusive palatali nel dialetto di San Giovanni in Fiore (CS), in *La fonetica sperimentale. Metodo e applicazioni* (L. Romito, V. Galatà & R. Lio, editors), Torriana: EDK, 349-371.

Morén, B. (2006), Consonant-vowel interactions in Serbian: Features, representations and constraint interactions, *Lingua,* 116, 1198-1244.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., Allik, J., Sinkkonen, J. & Alho, K. 1997), Language-specific phoneme representations revealed by electric and magnetic brain responses, *Nature,* 385, 432-434.

Näätänen, R., Pakarinen, S., Rinne, T. & Takegata, R. (2004), The mismatch negativity (MMN): towards the optimal paradigm, *Clinical Neurophysiology,* 115, 140-144.

Nenonen, S., Shestakova, A., Huotilainen, M. & Näätänen, R. (2005), Speech-sound duration processing in a second language is specific to phonetic categories, *Brain and Language*, 92*,* 26-32.

Recasens, D. & Espinosa, A. (2007), An electropalatographic and acoustic study of affricates and fricatives in two Catalan dialects, *Journal of the International Phonetic Association,* 37, 143-172.

Schmid, S. (2010), Les occlusives palatales du Vallader, in *Actes du XXVe Congrès International de Linguistique et Philologie Romanes* (P. Danler, editor), Tübingen: Niemeyer, 185-193.

Tervaniemi, M., Ilvonen, T., Sinkkonen, J., Kujala, A., Alho, K., Huotilainen, M. & Näätänen, R. (2000), Harmonic partials facilitate pitch discrimination in humans: electrophysiological and behavioral evidence, *Neuroscience Letters*, 279, 29-32.

Titova, N. & Näätänen, R. (2001), Preattentive voice discrimination by the human brain as indexed by the mismatch negativity, *Neuroscience Letters,* 308, 63-65.

Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., Czigler, I., Scepe, V., Ilmoniemi, R. J. & Näätänen, R. (1999), Brain responses reveal the learning of foreign language phonemes, *Psychophysiology,* 36, 638-642.

DOES A TALKER'S OWN RATE OF SPEECH AFFECT HIS/HER PERCEPTION OF OTHERS' SPEECH RATE?

Sandra Schwab
Université de Genève
sandra.schwab@unige.ch

1. ABSTRACT

Many studies have investigated the factors that may affect the perception of speech rate (e.g. Grosjean & Lane, 1976; Feldstein & Bond, 1981; Kohler, 1986; Greene, 1987; Crown & Feldstein, 1991), but very few studies have examined the role that the talker's own rate might play in his/her perception of others' rate. Among them, Lass & Cain (1972) investigated the hypothesis that a speaker's preferred rate depended on his actual rate. They indeed showed that speakers who produced slow rates preferred listening to slow rates, whereas fast speakers tended to prefer fast rates. This conclusion raises the question whether speech rate production affects not only speech rate preference, but also speech rate perception. To our knowledge, very few studies have tried to answer this question. Gósy (1991) formulated the hypothesis that "the speaker's own speech tempo determines his judgements concerning that of other people: the faster his own speech the less fast he perceives that of others" (p. 101). Gósy showed that speakers with different speech rates (very slow, slow, moderate, fast, very fast) did not perceive speech rate in a similar way. In the same direction, Koreman (2006) hypothesized that listeners' own speaking habits may affect their perception of speech rate. Nevertheless, his results failed to show an effect of the listener's rate on his/her perception of rate.

Considering the lack of totally conclusive results on the role that the talker's rate might play in rate perception, the objective of this research is to explore more deeply the hypothesis that speakers with different speech rates do not perceive speech rate in a similar way. To this end, we conducted a perception experiment.

In this experiment, participants were asked to listen to and estimate various samples at different speech rates (normal, fast and slow), using a magnitude-estimation task (Stevens, 1957). Results firstly showed a negative correlation between rate estimation and own rate at normal and slow rates (respectively, $r = -0.45$, $r = -0.39$, $p < 0.05$), but no correlation at fast rate ($r = -0.11$, ns): speakers with fast speech rate tended to under-estimate the sample speech rates (i.e. to give a lower numeric estimation) in comparison with slow speakers (at normal and slow rates). Secondly, and more interestingly, a regression analysis revealed that the own rate has a moderator effect on rate estimation, at all rates (normal: $t(781) = -5.67$, $p < 0.001$; fast: $t(781) = -2.06$, $p < 0.05$; slow: $t(781) = -6.46$, $p < 0.001$): the faster a listener speaks, the less his/her rate estimations raise as a function of heard rates, especially at normal and slow rates.

In sum, the present research has shown that the talker's rate plays a role in rate perception: fast speakers not only tend to under-estimate speech rate in comparison with slow speakers, but they are also less sensitive to rate changes.

2. INTRODUCTION

Speech rate – determined by articulation rate and by number and duration of pauses (see Grosjean & Deschamps, 1975 for a detailed description) – has been widely studied from various points of view for the past 50 years. Among the numerous studies, many have dealt with speech rate perception. The perception of speech rate refers to the metalinguistic activity that performs a listener when hearing a certain rate. In other words, speech rate perception corresponds to the impression a listener gets from the rate of his/her interlocutor. Research in this field has shown that the subjective rate estimation grows more quickly than the objective physical measurements, and that it rises in a non-linear way (Lane & Grosjean, 1973). Speech rate perception can indeed be described by Stevens' power function law (Cartwright & Lass, 1975), which assumes that sensation is proportional to the physical intensity raised to a given power (Stevens, 1957).

Among the researches on factors affecting the perception of speech rate, Grosjean & Lane (1976) showed that articulation rate was more important than pause time in speech rate perception. Acoustic-phonetic factors such as fundamental frequency (e.g. Feldstein & Bond, 1981; den Os, 1985), amplitude (e.g. Feldstein & Bond, 1981) and duration (e.g. Kohler, 1986) have been shown to also affect speech rate perception. Moreover, the influence of cognitive-linguistic variables such as canonical phonological structure (e.g. Koreman, 2006), language (e.g. Grosjean & Lass, 1977), task (e.g. Grosjean, 1978), and language pathology (e.g. Tjaden, 2000) has also been investigated in speech rate perception. Finally, studies have suggested that speech rate perception might vary as a function of extralinguistic factors, such as gender, relationship between speakers (e.g. Crown & Feldstein, 1991) and visual information (e.g. Greene, 1987).

Nevertheless, very few researches have studied the role that the talker's own rate might play in his/her perception of others' rate. For example, Lass & Cain (1972) investigated the hypothesis that a speaker's preferred speech rate depended on his actual speech rate. They showed a good correlation ($r = 0.61$) between speakers' preferred and actual speech rates: speakers who produced slow speech rates preferred listening to slow speech rates, whereas fast speakers tended to prefer fast speech rates. This conclusion raises the question whether speech rate production affects not only rate preference, but also rate perception. To our knowledge, very few studies have tried to answer this question. Gósy (1991) formulated the hypothesis that "the speaker's own speech tempo determines his judgments concerning that of other people: the faster his own speech the less fast he perceives that of others" (p. 101). She indeed showed that speakers with different speech rates (from very slow to very fast) did not perceive rate in a similar way. In the same direction, Koreman (2006) hypothesized that listeners' own speaking habits may affect their perception of speech rate. Nevertheless, his results failed to show an effect of the listener's rate on his rate perception.

Consequently, considering the lack of totally conclusive results on the role that the talker's rate might play in rate perception, the objective of this research is to explore more deeply the hypothesis that speakers with different speech rates do not perceive rate in a similar way. To this end, we conducted the perception experiment that is described below.

3. PERCEPTION EXPERIMENT

3.1 Method

3.1.1 Participants

Twenty-eight French speaking participants took part in this experiment. Their mean age was 27; 9 years.

3.1.2 Stimulus Materials

The stimulus materials used in this experiment consisted of naturally produced versions of a passage at normal, fast and slow rates, recorded by twenty-eight talkers. We used an actualized French version of the "Pop Fan Passage", which has been widely used in studies dealing with speech rate (Grosjean, 1972).

> "A vrai dire, je suis un jeune de quinze ans à peu près normal, ni un cas psychologique sérieux, ni un gars au-dessus des autres. J'écoute Graffiti FM, je coupe mes cheveux très court pour être à la mode, et je porte une boucle d'oreille, mais je ne pense pas être un véritable passionné de musique rap."

Recordings and measurements

Forty native French speakers (20 males and 20 females, mean age of 28 years) were recorded individually in a sound-treated booth. Talkers were instructed to read the passage at rates they considered as normal, fast and slow. Recordings began with three readings at normal rate and continued with three readings at slow rate. After a small break, talkers were asked to read once the passage at normal rate and then three times at fast rate. Each talker's speech was recorded via a microphone onto digital audio tape. As the first reading at each rate and the normal reading after the break served respectively as training and as recalibration, they were not included in the selection procedure.

We measured with *Praat* 3.8 (Boersma, 2001) the duration of speech and pauses for each two readings of each talker. We decided to consider a pause as a silent interval (sometimes with mouth noises and respiration) longer than 200ms, because we wanted to be able to distinguish real pauses from long stop consonant closures, especially at slow rates. Measurements criteria were the following: glottalization before a vowel, aspiration after a stop consonant, release schwa appearing at the end of some consonants and creaky voice were included in speech, whereas aspiration before speech, eventual sighs and mouth noises were included in pauses. From these measurements, we obtained the total speech time, the articulation time, the pause time, as well as the number of pauses.

As far as the syllable number was concerned, we identified and counted syllables in all productions on the basis of listening only. We made sure that our procedure was reliable by asking two judges to identify and count syllables in a subset of productions. As comparisons between judges showed similar syllable number and identification, we obtained speech rate (syll/min), articulation rate (syll/sec), as well as pause number and mean duration

(msec) for both readings at each rate for each subject.[1] Then, the mean across the two readings for each variable was computed for each talker, as well as the rate range (difference between slow and fast rates (syll/min)).

Participants and stimuli selection

Although the other temporal variables (articulation rate, pause number and duration) were also considered, the procedure selection was mainly based on the distribution of speech rate (syll/min). We chose a representative subset of productions according to the distributions of normal, fast and slow speech rates. We indeed selected the productions in such a way that their distributions at normal, fast and slow speech rates as well as the distribution of rate range (difference between slow and fast rates) matched as best as possible the entire set of productions. According to this criterion, we selected one production at each rate (normal, fast and slow) of 28 talkers (14 males and 14 females). We made sure that the differences between rate distributions of the entire set ($n = 40$) and rate distributions of the subset of productions ($n = 28$) were not significant (normal: $t(66) = 0.01$, ns.; fast: $t(66) = 0.28$, ns.; slow; $t(66) = 0.7$, ns.), nor was the difference between range significant ($t(66) = 0.3$, ns.).[2] Moreover, despite the overlap between the three speech rates (slow rate of some talkers corresponded sometimes to normal or even fast rate for other talkers, and inversely, fast rate of some talkers was closer to normal rate for others), statistical analyses showed not only a significant rate difference ($F(2, 54) = 243.83$, $p < 0.0001$), but also significant differences between each rate (Tukey HSD, $p < 0.01$).

The selection of the productions enabled us then to invite the 28 talkers who produced them to the perception experiment. In sum, according to the speech rate (normal, fast and slow) and the range distributions, we selected not only materials for the perception experiment – 28 productions at each of the rates (normal, fast and slow) –, but also the 28 talkers – 14 males and 14 females – who would participate in the perception experiment.

3.1.3 Procedure

The 84 selected productions (28 talkers x 3 rates) were split up into three parts (A, B and C), in such a way that one production of each talker appeared in each part, and that no more than two same rates (normal, fast or slow) followed each other. Moreover, we added one filler production at the beginning of each part, and three filler productions were chosen as practice.

Participants were run individually, or two by two. After listening to each production through headphones, they were instructed to perform a magnitude estimation task (see Stevens (1957) for details). In this task, the listener has to assign a number to each speech

[1] The important distinction between speech rate and articulation rate has to be kept in mind. The former refers to the number of units (e.g. words, syllables, phones) produced in a specific time, including pauses. The latter refers to the number of units expressed in a specific time, excluding pauses (Grosjean & Deschamps, 1975). Speech rate can be expressed in syll/min (Grosjean & Deschamps, 1975), in words/min (Goldman-Eisler, 1968) or in phones/sec (Gósy, 1991), while articulation rate is generally expressed in syll/sec (Grosjean & Deschamps, 1975) or in phones/sec (Koreman, 2006).

[2] We also made sure that the other temporal variables (articulation rate, pause number and duration) were similar between the 40 productions and the selected subset of productions. None of the differences was significant ($p > 0.14$).

rate he hears. The number 10 corresponds to what the listener considers a normal speech rate. Thus, the number he gives must be proportional to the normal speech rate (10). For example, 20 corresponds to a speech rate which is twice as fast as the normal speech rate, and 5 corresponds to a speech rate which is twice as slow as the normal speech rate.[3]

Each session consisted of the presentation of the three practice productions, followed by the presentation of the three parts (84 productions), with a break between each part. Half the participants heard the three parts in the order A, B and C, and the other half heard them in the reverse order (C, B and A).

3.1.4 Data analysis

Within each rate (normal, fast and slow) we collected the rate estimation of the 28 heard productions, given by the 28 participants (784 data for each rate). We also computed, within each rate, the mean estimation for each participant (28 data for each rate). By means of correlations and regression analysis, we explored the relationship between rate estimations given by the participants and their own speech rates. Remind that participants' own rates (normal, fast and slow, expressed in syll/min) were obtained thank to the readings they were instructed to do in the recording session described above.

3.2 Results and discussion

This experiment aimed at examining the relationship between speech rate production and perception using a magnitude estimation task. We first connected production and perception data by means of correlations. More precisely, we correlated participants' own rate and their mean rate estimation, separately for the normal, fast and slow rates. Remind that participants read the passage at normal, fast and slow rates and that they had to judge the rate of normal, fast and slow speech samples.

Figure 1 represents rate estimation as a function of own rate for normal, fast and slow rates. Rate estimation is presented in logarithmic values on the left y-axis, while the corresponding raw values are presented in the right y-axis. In the same way, own rate appears in logarithmic values on the lower x-axis, whereas the corresponding values in syll/min appear in the upper x-axis.

[3] Participants could use whatever number they wanted; they were not limited to the numbers 5, 10 and 20.

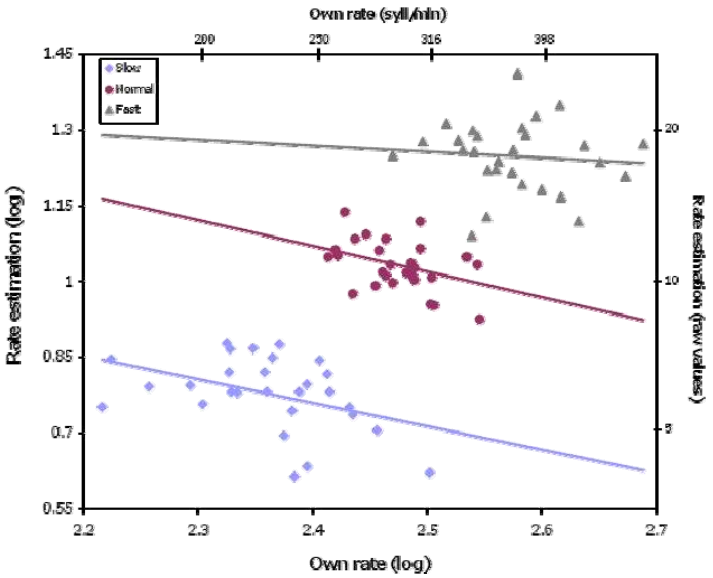


Figure 1: Rate estimation as a function of own rate for normal, fast and slow rates

Listeners are able to distinguish between the three rates: they give estimations that differ significantly between the rates (normal = 10.82, fast = 17.91, slow = 6.07; $F(2, 54) = 488.84$, $p < 0.001$)[4]. Consequently, it seems that the overlap we mentioned between the three rates in production doesn't impair the listeners' rate differentiation ability in the estimation task.

Secondly, and more interestingly, as can be seen in Figure 1, we find a negative correlation between rate estimation and own rate at normal and slow rates (respectively, $r = -0.45$, $r = -0.39$, $p < 0.05$), but no correlation at fast rate ($r = -0.11$, ns). These results show that speakers with fast speech rate tend to under-estimate the sample speech rates (i.e. (to give a lower estimation number) in comparison to slow speakers, especially at normal and slow rates

These results only partly confirm the hypothesis of a relationship between speech rate production and speech rate perception. This relation exists at normal and slow rates but not at fast rate. Further investigation is needed to examine in more details the reasons leading to the absence of a correlation between speech rate production and speech rate perception for fast rates. Indeed, fast speech rate might result in articulatory cues such as deletions (e.g. schwa deletions) or blurred speech, which might facilitate the perception of fast speech (Koreman, 2006), whatever the listener's own rate may be.

Further analyses were conducted in order to determine how the relationship between heard rate and rate estimations varied as a function of own rate. As illustrated in Figure 2, we were interested in studying whether the relationship between Heard rate (Independent

[4] Note that we used logarithmic values in statistic analyses, but for sake of clarity we present means in raw data.

variable, IV) and Rate estimation (Dependent variable, DV) varies according to Own rate, which might play the role of moderator.

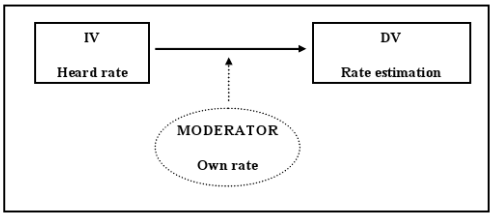


Figure 2: Relationship between Heard rate (IV) and Rate estimation (DV) according to Own rate (Moderator)

Given that moderator variables are characterized statistically in terms of interactions, we included in our regression model the interaction (i.e. cross-product term) of Heard rate and Own rate.[5] Therefore, we ran three separate regression analysis for the normal, fast and slow rates, with Rate estimation as a dependent variable, and Heard Rate, Own Rate and Interaction as independent variables. As the three independent variables (Heard rate, Own rate and Interaction) were highly correlated (VIF > 10),[6] the regression analyses were performed with Rate estimation as the dependent variable, and Heard rate and Interaction as independent variables (VIF = 2), separately for the normal, fast and slow rates.

As expected, regression analyses show first an effect of Heard rates, at all rates (normal: $t(781) = 21.05$, $p < 0.001$; fast: $t(781) = 19.25$, $p < 0.001$; slow: $t(781) = 24.01$, $p < 0.001$), meaning that Heard rate has a strong impact on Rate estimation. In other words, listeners are able to perceive rate differences within normal, fast and slow speech samples, respectively. Secondly and more interestingly, results show a significant interaction Heard rate x Own rate on Rate estimation, at all rates (normal: $t(781) = -5.67$, $p < 0.001$; fast: $t(781) = -2.06$, $p < 0.05$; slow: $t(781) = -6.46$, $p < 0.001$): the faster a listener speaks, the less his/her rate estimations raise as a function of heard rates.

Figure 3 shows the conditional slope of Rate estimation on Heard Rate as a function of own rate, for normal, fast and slow rates. In other words, the figure shows, on the y-axis, the conditional slopes relating Heard rate and Rate estimation (conditional on Own rate), and on the x-axis, the Own rate (in logarithmic values in the lower x-axis, and in syll/min in the upper x-axis).

[5] Interaction refers here to the cross-product term of Heard rates and Own rate. For example, if Heard rate = 2.54 (in log), and Own rate = 2.46 (in log), thus Interaction = 2.54 x 2.46 = 6.25.

[6] The Variance Inflation Factor (VIF) "provides an index of the amount that the variance of each regression coefficient is increased relative to a situation in which all of the predictor variables are uncorrelated. […] A commonly used rule of thumb is that any VIF of 10 or more provides evidence of serious multicollinearity involving the corresponding IV" (Cohen, Cohen, West & Aiken, 2003; 423).

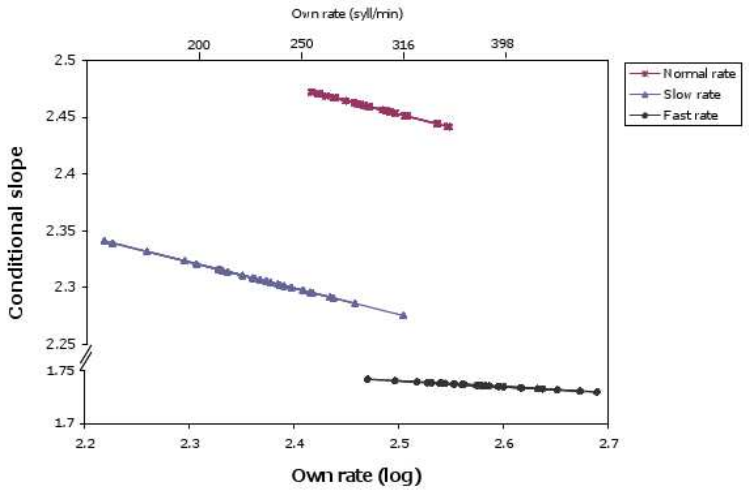


Figure 3: Conditional slope of Rate estimation on Heard Rate as a function of own rate, for normal, fast and slow rates

As can be seen, for all rates, the conditional slope decreases as a function of own rate. We can also observe that the slope is steeper at normal and slow rates than at fast rate, suggesting that own rate has a smaller effect on rate estimation at fast rate. In sum, own rate plays a moderator role in rate estimation: listeners with different speech rates perceive rate in a different way. The faster the own rate is, the less the rate estimation rises as a function of heard rates, especially at normal and slow rates.

4. CONCLUSION

The hypothesis we explored in this research was that speakers with different rates do not perceive speech rate in a similar way. More specifically, we hypothesized that fast speakers tend to under-estimate speech rate (i.e. to give a lower numeric estimation) in comparison with slow speakers. On one hand, participants were asked to read a passage at normal, fast and slow rates, and on the other hand, they were instructed to listen to and estimate various speech samples produced at different speech rates (normal, fast and slow), using a magnitude-estimation task.

Correlation analyses showed that speakers with fast speech rate tend to under-estimate the sample speech rates (i.e. to give a lower numeric estimation) in comparison to slow speakers (at normal and slow rates). Furthermore, regression analyses, which examined the moderator effect of own rate on rate perception, revealed that the faster the own rate is, the less the estimation rises as a function of heard rates. Therefore, the correlations between own rate and estimation, on one hand and, on the other hand, the moderator effect of own rate on rate perception suggest the existence of a relationship between speech rate production and perception, the former defining the latter.

As far as fast rate is concerned, further investigation is needed to study deeper the weakness of the relationship. Indeed, it might be due to a ceiling effect, more specifically to the fact that fast speech is easy to identify in presence of eventual blurred speech or

deletions, which would explain why fast speech is perceived as fast by all participants, whatever their own rate may be.

A question that may arise from these results concerns the direction of the relationship between speech rate production and perception. At the segmental level, the direction of the link between production and perception has been considered in both ways. Indeed, following the hypothesis of Perkell *et al.* (2004), speech perception affects speech production, while according to other researchers (Paliwal, Lindsay & Ainsworth, 1983) and defenders of the *Motor Theory of Speech Perception* (Liberman & Mattingly, 1985), speech production regulates speech perception. Following Gósy (1991) and Koreman (2006), we hypothesized that rate production regulates rate perception, but it would be worth considering the reverse possible interpretation.

In sum, we can conclude that a talker's own rate of speech does affect his/her perception of others' speech rate. More specifically, fast speakers not only tend to under-estimate (i.e. to give a lower numeric estimation) speech rate in comparison with slow speakers, but they are also less sensitive to rate changes. This finding highlights the importance of considering and controlling the listeners' own rate in experiments dealing with speech rate perception.

ACKNOWLEDGEMENTS

The author thanks Prof. François Grosjean (University of Neuchâtel, Switzerland) for his help in designing the experiment, and Prof. Boris Wernli (University of Neuchâtel, Switzerland) for his help in statistics.

5. REFERENCES

Boersma, P. (2001), PRAAT, a system for doing phonetics by computer, *Glot International*, 5, 341-345.

Cartwright, L.R. & Lass, N.J. (1975), A psychophysical study of rate of continuous speech stimuli by means of direct magnitude estimation scaling, *Language and Speech*, 18, 358-365.

Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003), *Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences* (Third Edition), Mahwah, NJ: Lawrence Erlbaum Associates.

Crown, C.L. & Feldstein, S. (1991), The perception of speech rate from the sound-silence patterns of monologues, *Journal of Psycholinguistic Research*, 20, 47-63.

den Os, E. (1985), Perception of speech rate of Dutch and Italian utterances, *Phonetica*, 4, 124-134.

Feldstein, S. & Bond, R. (1981), Perception of speech rate as a function of vocal intensity and frequency, *Language and Speech*, 24, 387-395.

Goldman-Eisler, F. (1968), *Psycholinguistics: Experiments in Spontaneous Speech*, London: Academic Press.

Gósy, M. (1991), The perception of tempo, in *Temporal Factors in Speech. A collection of Papers* (M. Gósy, editor), Budapest: Research Institute for Linguistics, HAS, 63-107.

Green, K.P. (1987), The perception of speaking rate using visual information from a talker's face, *Perception & Psychophysics*, 42, 587-593.

Grosjean, F. (1972), *Le rôle joué par trois variables temporelles dans la compréhension orale de l'anglais étudié comme seconde langue, et perception de la vitesse de lecture par des lecteurs et des auditeurs*, Thèse de Doctorat, Université de Paris VII, Paris, France.

Grosjean, F. (1978), *Perception of rate and processing of sentences*, Unpublished Manuscript, Northeastern University, Boston, USA.

Grosjean, F. & Deschamps, A. (1975), Analyse contrastive des variables temporelles de l'anglais et du français: Vitesse de parole et variables composantes, phénomènes d'hésitation, *Phonetica*, 31, 144-184.

Grosjean, F. & Lane, H. (1976), How the listener integrates the components of speaking rate, *Journal of Experimental Psychology: Human Perception and Performance*, 2, 538-543.

Grosjean, F. & Lass, N. (1977), Some factors affecting the perception of reading rate in English and in French, *Language and Speech*, 20, 198-208.

Kohler, K.J. (1986), Parameters of speech rate perception in German words and sentences: Duration, F0 movement, and F0 level, *Language and Speech*, 49, 115-139.

Koreman, J. (2006), Perceived speech rate: The effect of articulation rate and speaking style in spontaneous speech, *Journal of the Acoustical Society of America*, 119, 582-596.

Lane, H. & Grosjean, F. (1973), Perception of reading rate by speakers and listeners, *Journal of Experimental Psychology*, 97, 141-147.

Lass, N.J. & Cain, C.J. (1972), A correlational study of listening rate preferences and listeners' oral reading rates, *Journal of Auditory Research*, 12, 308-312.

Liberman, A.M. & Mattingly, I.G. (1985), The motor theory of speech perception revised, *Cognition*, 21, 1-36.

Paliwal, K.K., Lindsay, D. & Ainsworth, W.A. (1983), Correlation between production and perception of English vowels, *Journal of Phonetics*, 11, 77-83.

Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Stockmann, E., Tiede, M. & Zandipour, M. (2004), The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts, *Journal of the Acoustical Society of America*, 116, 2338-2344.

Stevens, S.S. (1957), On the psychophysical law, *Psychological Review*, 64, 153-181.

Tjaden, C. (2000), A preliminary study of factors influencing perception of articulatory rate in Parkinson disease, *Journal of Speech, Language and Hearing Research*, 43, 997-1010.

CROSS-LANGUAGE SPEECH PERCEPTION: LEXICAL STRESS IN SPANISH WITH ITALIAN AND FRANCOPHONE SUBJECTS[1]

Iolanda Alfano [a], Sandra Schwab [b], Renata Savy [c], Joaquim Llisterri [a]
[a] Universitat Autònoma de Barcelona; [b] Université de Genève; [c] Università di Salerno
Iolanda.Alfano@campus.uab.cat, sandra.schwab@unige.ch, rsavy@unisa.it, Joaquim.Llisterri@uab.cat

1. ABSTRACT

The present work analyses the perception of lexical stress in Spanish by Italian and French native speakers, trying to take into account the differences between Italian, French and Spanish stress systems both in perception and in production.

We have designed two perception experiments using a similar procedure of other works with native subjects (Llisterri *et al*., 2005; Alfano, 2006) and non-native subjects (Alfano *et al*., 2007; Alfano *et al*., 2009). The corpus consisted of couples and triplets of meaningful three syllable words and meaningless three syllable words (pseudo-words), with three different possible stress patterns: proparoxytone (PP), paroxytone (P) and oxytone (O). It was analyzed using the Praat software (Boersma & Weenink, 2003); for each of the three vowels of the words, we measured fundamental frequency (f_0) and vowel duration (D). F_0 and D values of the original words were systematically manipulated and a resynthesis was performed to create the stimuli used in the test, so that the role of these cues in perception could be studied. The test stimuli were created in the following way: in proparoxytone words, f_0 and duration values for each vowel were replaced by the corresponding f_0 and duration values found in the equivalent paroxytone words (PP>P); in the same way, in P words, f_0 and duration values for each vowel were replaced by the corresponding f_0 and duration values found in the equivalent oxytone words (P>O).

Three groups of Italian subjects and two groups of Francophone subjects, divided according to their competence in Spanish, have participated in the experiments performing an identification test.

Italian subjects perceive correctly the stressed syllable in almost 100% of the cases for the original PP and P, but make a mistake in more than 15% of the cases for the original O. The isolated manipulation of D or f_0 does not trigger a change in stress pattern perception. When f_0 and duration values are simultaneously modified, subjects perceive a change in stress location in a high percentage of cases for PP>P (between 56,7% and 90%) but in less than 40% for P>O.

Francophone subjects recognize the stressed syllable in original stimuli in more than 70% of the cases, obtaining the best results in PP perception and the worst performance for O. Such as the previous case with Italian speakers, the isolated manipulation of one of the two acoustic cues does not produce a clear change in stress pattern perception, while the combined manipulation produces a change in stress perception up to 77% of the cases for PP>P and in percentages ranging from 32 to 44% of the cases for P>O.

[1] The work is a result of the collaboration among the authors; nevertheless, we owe to Iolanda Alfano the paragraphs 1 and 2, to Sandra Schwab the paragraph 4, to Renata Savy the paragraph 3 and to Joaquim Llisterri the paragraphs 5 and 6.

Our results indicate that native language influence is not sufficient to explain the perception process in a foreign language, suggesting that non-native subjects use not only purely 'linguistic' perception strategies, since their choices seem to be also determined by psycholinguistic factors and by acoustic properties of the signal.

2. INTRODUCTION

The study of speech perception in a foreign language has raised many unsolved issues related with several aspects. A large amount of research studies focus on how non-native speakers perceive foreign speech and try to find out the most important factors that influence this process, such as the transfer from the native language, the role played by the level of knowledge in a foreign language or the interdependence between perception and production skills in that language.

Several models, mostly concentrating on segmental feature perception, have been developed in order to predict the stages of the perception process. Moreover, experimental studies on suprasegmental features perception seem to indicate a strong influence of native language too.

2.1 Models of second/foreign language acquisition/learning

In order to predict the stages of acquisition/learning,[2] numerous models have been developed; we will very briefly summarize some of the most frequently discussed in the literature on L2 acquisition.

The *Perceptual Assimilation Model* (PAM; Best, 1993 and 1994) explains the differential sensitivity to foreign language contrasts by appealing to the notion of phonological perceptual assimilation. According to this model, there are three ways in which a non-native segment may be perceptually assimilated to the native phonological system: (1) mapped onto a native phoneme varying in range from an excellent to a poor exemplar, (2) uncategorized phone falling in between native categories (i.e., roughly equally similar to more than one phoneme) and (3) nonassimilable sound that is very different from any native phoneme and fails to be assimilated within native phonological space. This model suggests a strong influence of the first language on the perception of a second language, but does not seem to give a real prediction of the process stages.

Based on a similar idea of that of the PAM, the *Native Language Magnet* (NLM; Kuhl, 1991 and 2000) considers a perception space in which prototypic native sounds work as a magnet, since they attract L2 sounds that are perceptually similar. The model predicts that all the cases of similar L2 sounds may be problematic because they are difficult to discriminate.

This model shares his background with the *Speech Learning Model* (SLM; Flege, 1995). According to the *SLM*, the greater the perceived phonetic dissimilarity between an

[2] The two terms are not usually used with the same meaning. Acquisition (of a second language) is considered a subconscious process of which the individual is not aware; this process is similar to the process that children undergo when learning their native language, since it takes place in a natural life context. Learning a foreign language, on the other hand, is considered a conscious process, that involves some kind of formal instruction, with rules and grammar (Krashen, 1987). As a matter of fact, it is not difficult to imagine how this distinction can be sometimes impossible to apply, since there are so many intermediate cases. For a review on this topic, see Manchón Ruiz (1987).

L2 speech sound and the closest L1 sound is, the more likely learners will be to discern the difference between the L1 and L2 sounds and show measurable progress in production and/or perception. The initial disadvantage of the more dissimilar L2 speech sound will ultimately prove to be an advantage. By hypothesis, a relatively high degree of perceived dissimilarity will eventually result in accurate segmental production and perception because it will promote the formation of a new category.

These models make a connection between the two linguistic systems considered (the one of the L1 and the one of the L2), linguistic perception and phonological acquisition of a new category. Also the *Feature Competition Model* (Brown, 2000) considers that the most frequent phonological categories influence perception and categorization of new sounds. This model implies the idea of a perceptual assimilation process and proposes an algorithm to determine how a category is prominent in order to predict its influence.

The models we have briefly discussed share some ideas and agree on several factors that may influence speech perception. It is interesting to notice that even if they develop different theoretical frameworks, they all agree on the *influence* of the *L1* phonological structure on L2 sounds perception. Nevertheless, it has to be observed that these models focus only on segmental and not suprasegmental features.

2.2 L2 Lexical stress perception experiments

A large amount of research on non-native speech perception has focused its attention on segmental features (that L1 and L2 do not share), while considerably less attention has been paid to non-native perception of suprasegmentals; that is probably due to the difficulties involved in the analysis of suprasegmental features. Moreover, as far as we know, the studies on the perception of suprasegmental features by non-native subjects often consider typologically different languages (among others, Wang *et al.*, 1999).

Nonetheless, both research on segmental and on suprasegmental features suggest an influence of L1 characteristics on L2 perception (Flege & Hillenbrand, 1986; McAllister *et al*., 2002; Cutler *et al*., 1986; Otake *et al*., 1993).

As regards L2 lexical stress perception, a large amount of research has focussed on free stress *vs.* fixed stress languages, such as French *vs.* Spanish, while, as far as we know, less attention has been given to more closely related languages, as it is the case of Italian and Spanish.

Since Francophone speakers seem to be unable to produce some stress patterns, it has been hypothesised that they can be insensitive to stress differences or 'deaf' to stress. Many studies focus on how Francophone speakers perceive lexical stress in Spanish. Some of them mitigate the idea of a complete stress deafness (Mora *et al*., 1997; Muñoz *et al*., 2008); others indicate that the ability to perceive stress location depends on the degree of the cognitive charge of the task. Dupoux and colleagues, following different experimental procedures, suggest that subjects' sensibility to stress depends on the possibility to rely on acoustic cues, since Francophone speakers show some problems in difficult tasks with ABX paradigms, but do not result deaf to stress in easier tasks with AX paradigms (Dupoux *et al*., 1997, 2001 and 2008; Peperkamp *et al.*, 1999).

Finally, others works strongly suggest the need to take into account the acoustic characteristics of the signal, since perception closely depends on them, both in L1 and in L2 (Alfano *et al*., 2007, 2008, 2009; Wang, 2008).

2.3 Aims and hypothesis

The purpose of this research is to examine thoroughly:

- perceptual strategies assumed by non-native speakers, at least as far as lexical stress in isolated words is concerned;
- the influence of native language on perception;
- the role of the level of L2 knowledge (Spanish, in our case).

It is well known that Italian, Spanish and French share important properties – among them, they show a tendency to isosyllabicity[3] – but, as regards their accentual systems, we have to consider that both Italian and Spanish are free stress languages, while French is a fixed stress language. For this reason, since our initial hypothesis was that perceptual strategy closely depends on the native language, we expected that, listening to Spanish stimuli, native Italian subjects would behave quite differently from native French speakers.

3. EXPERIMENT WITH ITALIAN SUBJECTS

3.1 Subjects

Three groups of ten Italian speakers were tested individually. The first group (group A) had been studying Spanish for several months (6-7 months); the second one (group B) although had never studied Spanish, knew some Spanish thanks to travels to Spain or listening to Spanish music; the third one (group C) had never studied Spanish and had never had any kind of contact with this language.

The thirty Italian subjects, aged 17 to 54 years, were born in the Italian region of Campania and had been living there for many years. Moreover, subjects were selected by means of a preceding test: they were asked to perform a previous identification task, in order to be sure that there were able to correctly identify the stress location.

3.2 Corpus

As Table 1 shows, the corpus consisted of six couples of meaningful three syllable words (*words*) and six couples of meaningless three syllable words (*pseudo-words*).

words	
['baskula] - [bas'kula] 'Scales'-'He/she/it swings'	[der'riβo] - [derri'βo] 'I pull down'-'He/she/it pulled down'
['kantara] - [kan'tara] 'large pitcher (or a liquid measure)'-'He/she sang'	[re'traso] - [retra'so] 'delay'-'He/she/it was late'
['lastima] - [las'tima] 'pity'-' He/she/it hurts'	[bor'rara] - [borra'ra] 'He/she/it deleted'-' He/she/it will delete'
pseudo-words	
['maleðo] - [ma'leðo]	[ma'leðo] - [male'ðo]
['laðeβo] - [la'ðeβo]	[la'ðeβo] - [laðe'βo]
['nuliβo] - [nu'liβo]	[nu'liβo] - [nuli'βo]

Table 1: Corpus used with Italian subjects (proparoxytone, paroxytone and oxytone words and pseudo-words)

[3] The division between syllable-timed languages and stress-timed languages constitutes a vexed question (see, among others, Bertinetto, 1989 and 1990; Bertinetto & Magno Caldognetto, 1993; Almeida, 1997; Cantín & Ríos A., 1991; Ramus *et al*., 1999; Russo & Barry, 2008).

3.3 Method

The experimental procedure we have adopted had been designed for a study with native Spanish subjects (Llisterri *et al*., 2005) and already followed in other works with native Italian subjects (Alfano, 2006) and non-native subjects (Alfano *et al*., 2007; Alfano *et al*., 2009).

The corpus was read 10 times by a native Spanish speaker;[4] for each of the three vowels of the stimuli, we measured:

- f_0 at the beginning, at the centre and at the end of the vowel;
- vowel duration.

The corpus was analyzed and resynthesized using the Praat software (Boersma & Weenink, 2003).

The test stimuli were created in the following way: the original values of both f_0 and duration were replaced in each vowel of each stimulus by the mean values of each parameter, using Praat PSOLA algorithm (hereafter, *Original stimuli*). Moreover, in proparoxytone words, f_0 and duration values for each vowel were replaced by the corresponding f_0 and duration values found in the equivalent paroxytone words (PP>P *Manipulated stimuli*); in the same way, in P words, f_0 and duration values for each vowel were replaced by the corresponding f_0 and duration values found in the equivalent oxytone words (P>O *Manipulated stimuli*).

Each word was resynthesised with the replaced values using PSOLA as implemented in Praat; Figure 1 shows an example of a manipulated stimulus.

The values have been modified not only individually, but also simultaneously, obtaining the three possible combinations of *Manipulated stimuli*: f_0, D, f_0+D. This strategy has allowed the study of the effects of each acoustic cue both in isolation and in combination with the other.

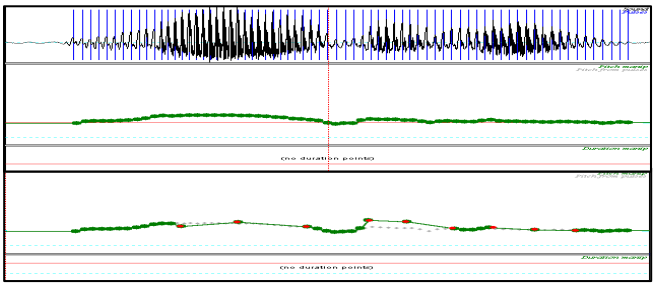


Figure 1: [’navilo] with the original f_0 contour (top) and after superimposing the f_0 contour of [na’vilo] (bottom)

3.4 Procedure

The tests were administered using a specifically designed data-collection software.[5] Subjects were told they were going to listen to Spanish words and pseudo-words; each participant received instructions about the task and did a brief training.

[4] The speaker did not receive particular instructions, he was asked to read following the indicated stress pattern.

The experiment consisted of an identification test of the stressed syllable: subjects had to click on the key 1 if they perceived stress location on the first syllable, on the key 2 if they thought the stressed syllable was the second one and on key 3 if they perceived stress location on the last syllable (see Figure 2).

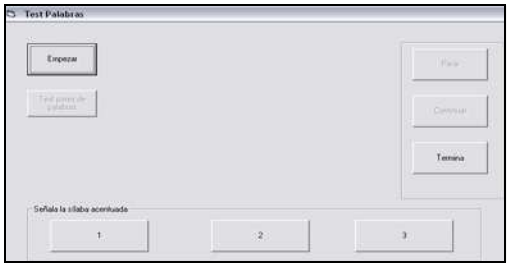


Figure 2: Screen of the identification test with Italian speakers

We proposed 57 stimuli (21 original stimuli plus 12 x 3 items with manipulated values-D; f_0; D+f_0). The stimuli were given in random order; a total of 1710 answers was obtained.

3.5 Results

3.5.1 Original stimuli

Figure 3 shows the results obtained for original stimuli, that is to say those items with no manipulation of the acoustic cues.

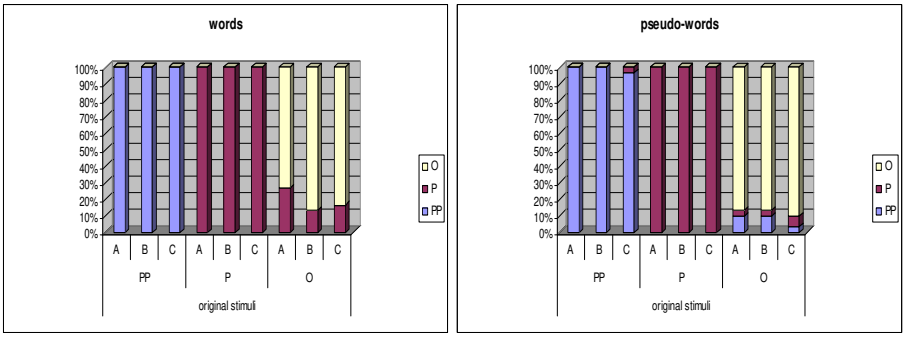


Figure 3: Results in % from identification test for original words (left) and pseudo-words (right); A, B, C = subject groups; PP = proparoxytones, P = paroxytone, O = oxytone

For original PP and P stimuli, the average of correct identification and discrimination reaches very high values, without relevant differences among the three groups and between words and pseudo-words. However, listening to original O subjects behave in a different way: group A does not correctly perceive the stress in 26,7% (words) and in 13,3% of the cases (pseudo-words); group B fails in 13,3% of the cases (words and pseudo-words) and group C in 16,7% (words) and in 10% of the cases (pseudo-words).

[5] The software has been designed by Dr. P. Riccardi (University of Naples).

3.5.2 Manipulation of f_0

Analysing the results of f_0 manipulation in words, it can be seen that PP>P stimuli are identified as paroxytone in percentages reaching the 33% of the cases for group A, the 10% for group B and the 26,7% for group C; looking at the answers concerning pseudo-words, it is possible to see that PP>P stimuli are perceived with a change in stress location (as P) in percentages ranging from 46,7% (group C) to 70% (groups A and B): words and pseudo-words answers do not show the same trend.

On the contrary, P>O stimuli are perceived as paroxytone, that is to say with the original stress pattern, in very high percentages, both in words and in pseudo-words (see Figure 4).

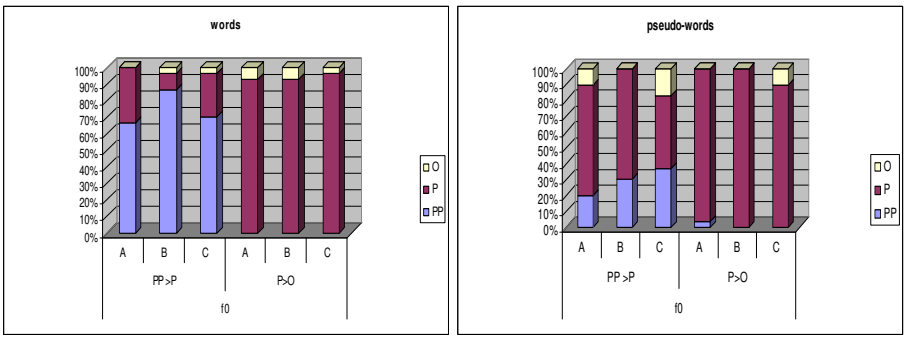


Figure 4: Results in % from identification test for words (left) and pseudo-words (right) with a manipulation of f_0 values; A, B, C = subject groups; PP = proparoxytones, P = paroxytone, O = oxytone; PP>P = proparoxytones with paroxytone f_0 values; P>O = paroxytone with oxytone f_0 values

3.5.3 Manipulation of duration

Looking at the answers concerning stimuli with modified duration (see Figure 5), it can be seen that subjects clearly perceive P>O stimuli as paroxytone, that is to say with the original stress pattern, both in words (83,3%, 94%, 86,7%, group A, B and C respectively) and in pseudo-words (93,4%, 90%, 96,7% , group A, B and C respectively).

As regards PP>P stimuli too, the manipulation of duration does not trigger a clear change in stress pattern perception, but it should be noted that group A behaves in a different way in the case of pseudo-words: it identifies PP>P pseudowords as paroxytone in 66,7% of the cases, while groups B and C identify pseudowords PP>P as paroxytone in only 3,3% and 16,7% of the cases (see Figure 5).

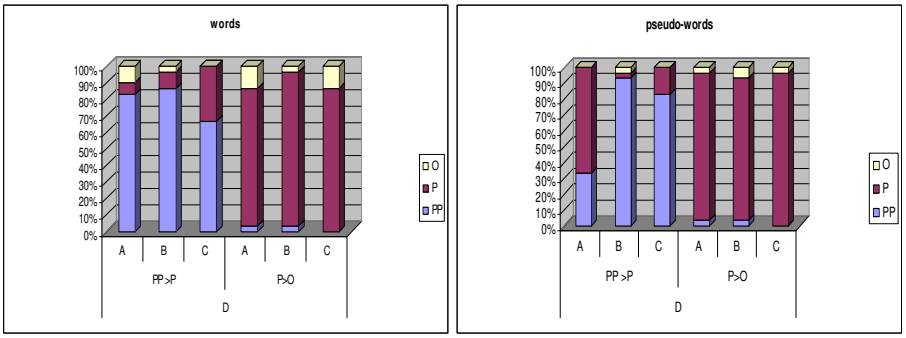


Figure 5: Results in % from identification test for words (left) and pseudo-words (right) with a manipulation of duration values; A, B, C = subject groups; PP = proparoxytones, P = paroxytone, O = oxytone; PP>P = proparoxytones with paroxytone duration values; P>O = paroxytone with oxytone duration values

3.5.4 Simultaneous manipulation of f_0 and duration

When f_0 and duration values are simultaneously modified, subjects perceive a change in stress location in a higher percentage of cases, both in words and pseudo-words, but it is interesting to notice that the manipulation has a stronger effect in the case of PP>P stimuli in comparison with P>O items (67,8% of the cases in favour of P words , but only 28,9% in favour of O words, respectively on the left and on the right side of each graphic of Figure 6).

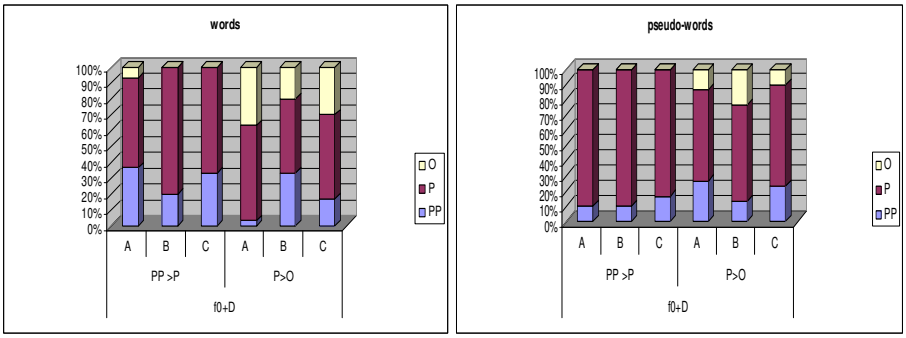


Figure 6: Results in % from identification test for words (left) and pseudo-words (right) with a manipulation of f_0 and duration values; A, B, C = subject groups; PP = proparoxytones, P = paroxytone, O = oxytone; PP>P = proparoxytones with paroxytone f_0 and duration values; P>O = paroxytone with oxytone f_0 and duration values

3.5.5 Remarks

It is evident that Italian subjects show some problems when they listen to Spanish stimuli - even in the case of original items - depending on the stress pattern: they correctly perceive PP and P patterns, but make mistakes in identifying the oxytone pattern, both in words and pseudo-words (see Figure 3).

When f_0 values are modified, subjects seem to be indifferent to the manipulation in the case of P>O stimuli, but tend to perceive some change in stress pattern in PP>P, especially for pseudo-words (see Figure 4): Italian subjects do not react in the same way they do in their native language, since exposed to Italian items they never perceive the manipulation of f_0 (Alfano, 2006).

On the other hand, Italian subjects did perceive changes in stress location in the case of Italian stimuli with manipulation of duration, while they do not seem to be aware of duration differences in Spanish stimuli (except the case of the group A for pseudo-words, see Figure 5). Compared to Spanish stimuli with manipulated duration, Italian subjects do not behave in the same way they do with stimuli in their own language but, at the same time, do not rely on the acoustic cues used by native Spanish speakers (Llisterri *et al.*, 2005).

In the case of the simultaneous manipulation of both acoustic cues, the percentages of answers that indicate a change in stress perception are the 67,8% for PP>P, but only the 28,9% for P>O,[6] where in Italian L1 they were respectively the 90,8% and the 71,7% (Alfano, 2006).

First of all, to interpret the results, it is important to consider that the higher frequency of paroxytone words in Spanish and Italian may bias the processing of oxytone words towards the most common pattern:[7] in the case of PP>P stimuli, subjects may be more sensitive to the manipulation in comparison with P>O items, since it goes in the direction of paroxytone pattern. However, problems with oxytones have been also detected in the case of original items.

Secondly, we think that differences of acoustic duration between Spanish and Italian stressed vowels could be considered as one of the reasons of this behaviour, especially with oxytone words. The acoustic analysis of the stimuli shows that in internal word position Italian stressed vowels are 35,8% longer than Spanish ones, but in oxytone words they are 12% shorter. Moreover, prepausal stressed vowels are, in Spanish, 42,3% longer than word internal ones, while in Italian they are 7,8% shorter than word internal ones (Alfano *et al*., 2009). For this reason, Italian subjects seem to be somehow unable to solve a sort of conflict between the acoustic stimuli (that is to say long final stressed vowels) and their L1 expectations (short final stressed vowels).

Comparing pseudo-words with words results, further investigation is needed not only to understand better the difference observed between words and pseudowords in PP>P f_0 manipulated stimuli, but also to explore more deeply the differences between the three groups in PP>P pseudowords manipulated in duration.

We have carried out the experiment on three different groups of subjects in order to point out possible differences depending on the level of Spanish knowledge. The analysis of each group does not reveal a clear trend: it can be observed that, in some cases, group A seems to behave in a quite different way from the other ones. Nevertheless, we need to stress the fact that lexical knowledge does not seem to constitute a very important factor, since group A knew the meaning of the items in the only the 25% of the cases and group B

[6] Average percentages of the three groups, concerning words stimuli (§ *3.5.4*).

[7] For a comparison between Italian and Spanish systems, concerning the distribution of the different stress patterns, see Alfano (2008).

did not reach the 10%.[8] In any case, we must go carefully and consider that the three groups did not have a very different level of Spanish knowledge (§ 3.1).

In sum, we can conclude that the performance of non-native Italian subjects appears to be influenced by their native language but, at the same time and in a strong way, by the acoustic features of the signal too.

4. EXPERIMENT WITH FRENCH SPEAKERS

4.1 Subjects

Two groups of French speaking subjects took part in this experiment: one French speaking group with advanced knowledge in Spanish and one French speaking group with no knowledge in Spanish. The advanced group in Spanish (hereafter, group A) was composed of 10 subjects. They were between 21 and 36 years old and were all raised in a French speaking environment with only one language, French. They had been studying Spanish at University of Neuchâtel (Switzerland) during 6-11 years.

As far as the French speaking group with no knowledge in Spanish (hereafter, group B) is concerned, it was composed of 10 students of the University of Neuchâtel. They were between 19 and 24 years old and were all raised in a French speaking environment with only one language, French. Although some of these subjects indicated good knowledge in German and/or English (learning these two languages is obligatory in the Swiss educational system), none of them reported good knowledge in Italian (which excludes the eventual bias of knowing a free stress Romance language).

4.2 Corpus

The corpus we used was taken from Llisterri *et al*. (2005). It was composed of 4 triplets of trisyllabic words (CVCVCV) and 4 triplets of trisyllabic pseudo-words (see Table 2). All words and pseudo-words could be proparoxytones, (e.g. *número*), paroxytones (e.g. *numero*) and oxytones (e.g. *numeró*).

words
[’limite] - [li’mite] - [limi’te] ‘limit’ - ‘I limit’ - ‘I limited’
[’meðiko] - [me’ðiko] - [meði’ko] ‘doctor’-‘I medicate’ - ‘He/she medicated’
[’numero] - [nu’mero] - [nume’ro] ‘number’ -‘I number’ - ‘He/she numbered’
[’baliðo] - [ba’liðo] - [bali’ðo] ‘valid’ -‘I validate - ‘He/she validated’
pseudo-words
[’maleðo] - [ma’leðo] - [male’ðo]
[’laðeβo] - [la’ðeβo] - [laðe’βo]
[’nuliβo] - [nu’liβo] - [nuli’βo]
[’luɤiðo] - [lu’ɤiðo] - [luɤi’ðo]

Table 2: Corpus used with Francophone subjects (proparoxytone, paroxytone and oxytone words and pseudo-words)

[8] After taking the test, we asked the subjects to tell if they knew the meaning of the stimuli.

4.3 Method

The experimental procedure, which is described in detail in Llisterri *et al*. (2005), was similar to the one we used with Italian speakers (§ 3.3). In total, 24 *Original stimuli* and 48 *Manipulated stimuli* (16 x 3; D; f_0; D+f_0) were presented in this experiment.

4.4 Procedure

Subjects performed an identification task and were run individually. The stimuli were presented online from a laptop using DMDX software,[9] which also recorded the subjects' responses. Subjects were instructed to listen to each stimulus (e.g. *médico*), to make a selection among the three possible choices that appeared in a row on the computer screen (see Figure 7), and to press the corresponding button in a response box.

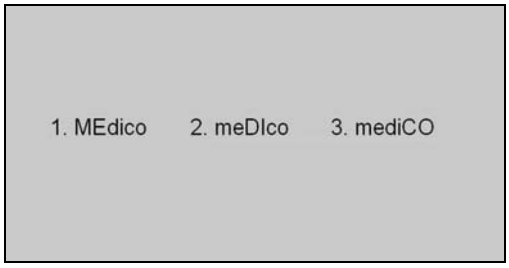


Figure 7: Screen of the identification test with French speakers

The left-to right order of the three choices was always the same across trials: Position 1 corresponded to the stimulus with stress on the first syllable, position 2 to the stimulus with stress on the second syllable, and position 3 to the stimulus with stress on the third syllable. Thus, subjects pressed button 1 when they perceived stress on the first syllable, button 2, for stress on the second syllable, and button 3 for stress on the third syllable. Each subject received a different randomization of the stimuli.

4.5 Results

4.5.1 Original stimuli

Figure 8 shows the results obtained for original stimuli, that is to say those stimuli with no manipulation of the acoustic cues. First of all, we observe that French speakers are able to correctly identify the location of stress in 71.5% of the cases (mean across patterns (PP, P and O), and across lexical status (words and pseudo-words)). Secondly, it appears that group A (advanced in Spanish) achieves a better performance than group B (with no knowledge in Spanish), whatever the pattern and the lexical status of the stimuli may be (mean correct identification (across patterns and lexical status) for group A = 82.2%; for group B = 60.8%).

[9] The software has been developed by K. Forster, Psychology Department (University of Arizona).

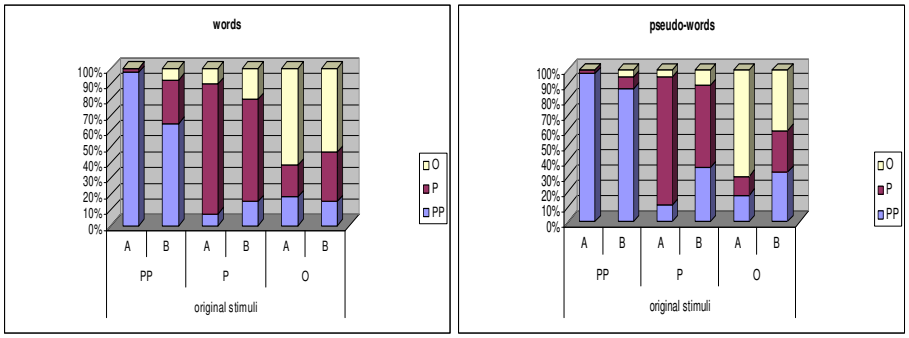


Figure 8: Results in % for original words (left) and pseudo-words (right); A, B = subject groups; PP = proparoxytones, P = paroxytone, O = oxytone

Thirdly, we notice that stress on the first syllable (PP) is better perceived (mean across groups and lexical status) = 86.7%) than stress on the second syllable (P; mean across groups and lexical status) = 71.5%), that is in turn better identified than stress on the third syllable (O; mean across groups and lexical status = 56.2%).

4.5.2 Manipulation of f0

Figure 9 shows the results obtained for stimuli with a manipulation of f_0 values. When f_0 is manipulated, French speakers perceive the change in stress pattern in 26.9% of the cases. As far as PP>P words are concerned, there are identified as paroxytone in percentages reaching the 34.2% of the cases for group A and the 27.5% for group B, while the difference between both groups increases with PP>P pseudo-words (group A = 37.5%; group B = 17.5%).

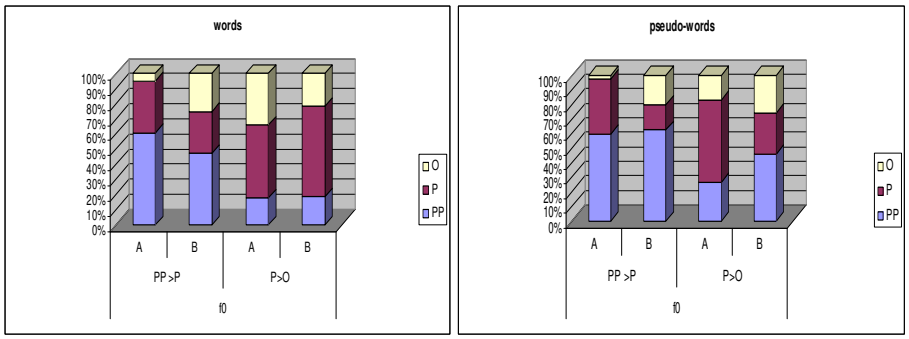


Figure 9: Results in % for words (left) and pseudo-words (right) with a manipulation of f_0 values; A, B = subject groups; PP = proparoxytones, P = paroxytone, O = oxytone; PP>P = proparoxytones with paroxytone f_0 values; P>O = paroxytone with oxytone f_0 values

Regarding P>O words, they are identified as oxytone in percentages reaching the 34.2% of the cases for group A, the 21.7% for group B, whereas the difference between both groups goes in the reverse direction with P>O pseudo-words (group A = 16.7%; group B = 25.8%). When we compare stress perception in PP>P and P>O stimuli, it appears that

the manipulation of f_0 seems to affect slightly more the perception of a change in stress pattern in PP>P stimuli (28.9%) than in P>O stimuli (24.5%).

4.5.3 Manipulation of duration

Figure 10 shows the results obtained for stimuli with a manipulation of duration values. French speakers perceive the change in stress pattern in only 13.1% of the cases. It is interesting to note that group B is more sensitive to the manipulation of duration than group A, whatever the pattern and the lexical status may be. Indeed, group B perceives the change in stress pattern in 18.5%, while group A perceives it in only 7.5% (means across patterns and lexical status). Moreover, the comparison of stress perception in PP>P and P>O stimuli shows that the manipulation of duration has a similar effect on both patterns (PP>P = 12.7%; P>O = 13.5%).

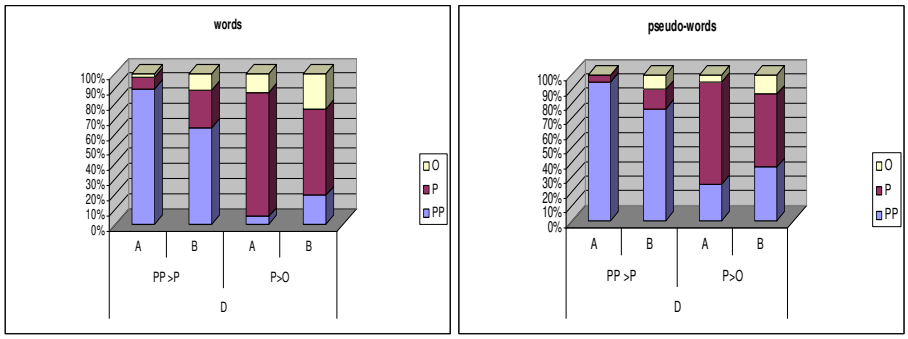


Figure 10: Results in % for words (left) and pseudo-words (right) with a manipulation of duration values; A, B = subject groups; PP = proparoxytones, P = paroxytone, O = oxytone; PP>P = proparoxytones with paroxytone duration values; P>O = paroxytone with oxytone duration values

4.5.4 Simultaneous manipulation of f0 and duration

Figure 11 shows the results obtained for stimuli with a manipulation of f_0 and duration values. The simultaneous manipulation of f_0 and duration triggers the perception of a change in stress pattern in 47.7% of the cases. Words and pseudo-words show the same trends. Firstly, the difference between group A and group B is similar whether it be words (means across patterns; group A = 60.8%; group B = 53.3%) or pseudo-words (means across patterns; group A = 41.7%; group B = 35.0%). Secondly, the simultaneous manipulation of f_0 and duration has a stronger effect on the perception of the accentual change in PP>P stimuli (57.5% across groups and lexical status) in comparison with P>O stimuli (37.9% across groups and lexical status).

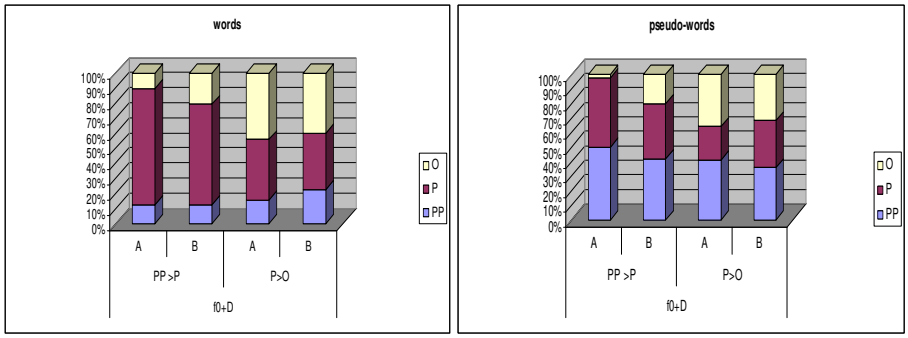


Figure 11: Results in % for words (left) and pseudo-words (right) with a manipulation of f_0 and duration values; A, B = subject groups; PP = proparoxytones, P = paroxytone, O = oxytone; PP>P = proparoxytones with paroxytone f_0 and duration values; P>O = paroxytone with oxytone f_0 and duration values

4.5.5 Remarks

In this experiment, French speakers had to identify the location of stress in stimuli with no acoustic manipulation (original stimuli) and in stimuli with separate and combined manipulation of f_0 and duration (manipulated stimuli). Results with original stimuli show first that French speakers are able to correctly perceive stress in 71% of the cases. This agrees with the results of Muñoz *et al*. (2008), who found a correct identification of 83% in a similar task, and indicates that French speakers might not be so deaf to stress as it was thought (at least in an identification task). Secondly, results suggest that the exposition to L2 makes French speakers more sensitive to stress, as the advanced group in Spanish identified stress more accurately than the group with no knowledge. Thirdly, it seems that it is harder for French speakers to perceive stress on oxytone stimuli (in original stimuli and in P>O stimuli). This observation, which was also highlighted in Muñoz *et al.* (2008), is quite surprising given the fact that French stress in mainly oxytone. It might be due to the different acoustic realization of stress in final syllables in French and Spanish. Nevertheless, more studies are needed to understand better the difficulty of French speakers to identify oxytone stress in L2.

As far as manipulated stimuli are concerned, results reveal first, as observed with Italian speakers, that the combined manipulation of f_0 and duration leads to a better perception of the accentual change than the separate manipulation of each acoustic parameter. It appears thus that stress is perceptually not defined by only one parameter, but by the combination of all parameters.

Secondly, results suggest that f_0 is a more important cue than duration for a syllable to be perceived as stressed by French speakers. Indeed, researches in French (Rigault, 1962; Dahan & Bernard, 1996) have shown that f_0 is the decisive parameter in the perception of prominences in French L1. It seems thus that French speakers have transferred this knowledge from L1 (French) to L2 (Spanish).

Thirdly, and more interestingly, both groups of French speakers (Advanced and With no knowledge) don't behave in the same way according to the different acoustic manipulations. On one hand, the advanced group perceives better the accentual change when both parameters (f_0 and duration) are jointly manipulated. On the other hand, while both groups

are equally sensitive to the isolated manipulation of f_0, the group with no knowledge in Spanish is more sensitive to the isolated manipulation of duration. It appears thus that French speakers with no knowledge in L2 process stress in a more acoustic way. Indeed, whereas the advanced French speakers in Spanish can process stress using their linguistic knowledge of Spanish, those with no knowledge rely more on all available acoustic cues, even less salient. Consequently, the knowledge in L2 seems to modify the perceptual strategies used in identifying stress in L2.

In sum, this experiment shows that French speakers' stress perception in L2 is affected not only by the native language, but also by the knowledge in L2, the accentual pattern of the stimuli and the acoustic parameters used in the realization of stress.

5. DISCUSSION

Native Italian subjects exposed to Spanish stimuli do not behave in the same way they do in their L1 - they do not react to the manipulation of duration (§ 3.5.3) and they perceive, in some cases, a change in stress pattern in the stimuli with f_0 modification (§ 3.5.2) – but, at the same time, they do not come to rely on the same acoustic cues used by native Spanish subjects (Llisterri *et al*., 2005).

Looking at the different patterns, it is quite interesting to notice the particular reaction to oxytone stimuli: as far as original items are concerned, the correct identification rate is lower in comparison with other patterns (§ *3.5.1*); moreover, no manipulation is sufficient to trigger a change in stress pattern perception in the case of P>O stimuli, that is to say that nothing seems to succeed in obtaining an item perceived as oxytone.

We suggest to take into account the following aspects: on one hand, as we have already discussed (§ 3.5.5), a typical oxytone Italian item is different from an oxytone Spanish stimulus and, on the other hand, the oxytone pattern is not very frequent in Spanish and absolutely uncommon in Italian,[10] so, in the case of PP>P stimuli, subjects may be more sensitive to the manipulation in comparison with P>O items, since it goes in the direction of the unmarked paroxytone pattern.

We believe that differences of acoustic duration between Spanish and Italian stressed vowels can help to understand the results, especially with oxytone words. The acoustic analysis of the stimuli shows that in internal word position Italian stressed vowels are 35,8% longer than Spanish ones, but in oxytone words they are 12% shorter. Moreover, prepausal stressed vowels are, in Spanish, 42,3% longer than word internal ones, while in Italian they are 7,8% shorter than word internal ones (Alfano *et al*., 2009). For this reason, we think Italian subjects are somehow unable to solve a sort of conflict between the acoustic stimuli (long final stressed vowels) and their L1 expectations (short final stressed vowels).

French speaking subjects are able to correctly perceive stress in 71% of the cases in stimuli with no acoustic manipulation: it means that French speakers might not be so deaf to stress as it was thought (at least in an identification task). As far as manipulated stimuli

[10] As far stress patterns distribution is concerned, Spanish and Italian do not show the same distributions: Spanish presents more oxytone words than Italian and in Italian we find more proparoxytones than in Spanish, but it is interesting to take into account the higher frequency of paroxytone words in both languages (see Alfano, 2008).

are concerned, the manipulation of f_0 seems to be necessary, but not sufficient, to determine a change in stress pattern perception.

Analysing the different patterns, it is unexpected that, as observed with Italian speakers, also French speaking subjects obtain the worst performance exposed to oxytone stimuli and P>O items. In the first place, we consider it might be due to the different acoustic realization of stress in final syllables in French and Spanish. However, further investigation is required to understand better the consequences of these differences on the perception in L2: a contrastive acoustic analysis between French and Spanish oxytones will probably allow to formulate a more solid hypothesis. In the second place, it is quite interesting to notice that the difficulty of French speakers with oxytone stress pattern in L2 is documented in other studies (Muñoz *et al.*, 2008), in which the authors suggest to take into account the so-called 'law of the grasp of consciousness'. This Claparède's law is based on the idea that the more often a sort of behavior or judgment has been used automatically or by habit, the harder it is to become aware of it. Also this factor could help to explain the reason why French speakers obtain the worst performance exposed to oxytone stimuli, since, as it is well known, French almost consistently stresses the last syllable. On one hand, therefore, French speakers could not be well aware of their native-language stress pattern and so they could not be able to easily recognize it and, on the other hand, it is possible that they do not pay enough attention when they are asked to identify oxytone stimuli. Finally, we can hypothesize that French speakers could be somehow aware of the differences between French and Spanish lexical stress and that this sort of awareness could give place to a form of 'perceptive hypercorrection', thwarting the identification of oxytone pattern. With no doubt the special status of oxytone pattern constitutes a crucial point and deserves a detailed analysis from each point of view we have only briefly discussed.

In both experiments we have considered different groups of subjects depending on their level of foreign language knowledge. While in the first experiment, with three groups of Italian speakers, the analysis of each group does not reveal a clear trend – even if it can be observed that, in some cases, group A (the most advanced subjects in Spanish) seems to behave in a quite different way from the other ones -, in the second experiment, with one group of French speaking subjects with advanced knowledge in Spanish (group A) and one French speaking group with no knowledge in Spanish (group B), we can see that group B seems to process stress in a more acoustic way in comparison with group A. It can be seen, therefore, that in the second experiment the knowledge in L2 seems to modify the perceptual strategies used in identifying stress in L2. Nevertheless, we have to take into account the fact that the three Italian groups did not have a very different level of Spanish knowledge (§ 3.1) and it is probably for this reason that we cannot evaluate in a deep way the role played by L2 knowledge.

6. CONCLUSIONS

In conclusion, our results clearly indicate that the perception of lexical stress in L2 is not an easy task - even in the case of a free stress language speakers - and confirm that it does not depend on only one acoustic parameter, but, such as in L1, it always depends on

the co-variation of both parameters (f_0 and duration).[11] Moreover, the obtained data indicate that lexical stress perception in L2 is affected not only by the native language, but also by the level of L2 knowledge, the stress pattern of the proposed items and the acoustic features of the signal.

It seems to be crucial, indeed, to consider the acoustic parameters used in the realization of stress in L1 and L2, since, as we have seen, the differences in the temporal organization contribute to create the expectations biased by the speakers' L1, and, therefore, have an influence on the perception in L2.

Nevertheless, further research is needed: it will be necessary to investigate better the factor of the level of L2 knowledge and to perform new acoustic analyses in the three languages in order to evaluate in a deeper way the importance of the acoustic features of the signal.

7. BIBLIOGRAPHY

Alfano, I. (2006), La percezione dell'accento lessicale: un test sull'italiano a confronto con lo spagnolo, in *Analisi prosodica. Teorie, modelli e sistemi di annotazione,* Atti del 2° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Salerno, 30 novembre-2 dicembre 2005 (R. Savy & C. Crocco, editors), Torriana: EDK Editore, 632-656.

Alfano, I. (2008), Strutture sillabiche ed accentuali in italiano e in spagnolo, in *Testi e linguaggi* (M. Voghera, editor), Roma: Carocci, 214-242.

Alfano, I., Llisterri, J. & Savy, R. (2007), The perception of Italian and Spanish lexical stress: A first cross-linguistic study, in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, August 6-10, 2007 (J. Trouvain & W.J. Barry, editors), 1793-1796.

Alfano, I., Savy, R., & Llisterri, J. (2008), Las características acústicas y perceptivas del acento léxico en español y en italiano: Los patrones acentuales paroxítonos, *Language Design. Journal of Theoretical and Experimental Linguistics. Special Issue 2: Experimental Prosody*, 2, 23-30.

Alfano, I., Savy, R., & Llisterri, J. (2009), Sulla realtà acustica dell'accento lessicale in italiano ed in spagnolo: la durata vocalica in produzione e percezione, in *La fonetica sperimentale. Metodo e applicazioni* (L. Romito, V. Galatà & R. Lio, editors), Atti del 4° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Arcavacata di Rende (CS), 3-5 Dicembre, 2007, Torriana: EDK Editore, 22-39.

Almeida, M. (1997), Organización temporal del español: el principio de isocronía, *Revista de Filología Románica*, 14, 29-40.

Bertinetto, P.M. (1989), Reflections on the dichotomy 'stress' vs. 'syllable-timing, *Revue de Phonétique Appliquée*, 91/93, 99-130.

[11] In this study we have examined the effects of f_0 and duration manipulation, but, as it is well known, the perception of lexical stress depends on the co-variation of three acoustic parameters: fundamental frequency, duration and also intensity.

Bertinetto, P.M. (1990), Coarticolazione e ritmo nelle lingue naturali, *Rivista Italiana di Acustica*, 14, 69-74.

Bertinetto, P.M. & Magno Caldognetto E. (1993), Ritmo e intonazione, in *Introduzione all'italiano contemporaneo. Le strutture* (A.A Sobrero, editor), Roma-Bari: Laterza, 141-192.

Best, C. (1993), Emergence of language-specific constraints in perception of non-native speech: A window on early phonological development, in *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* (B. de Boysson-Bardies, editor), Dordrecht: Kluwer Academic Publishers, 289- 304.

Best, C. (1994), The emergence of native-language phonological influence in infants: A perceptual assimilation model, in *The Transition from Speech Sounds to Spoken Words: The Development of Speech Perception* (H. Nusbaum, J. Goodman & C. Howard, editors), Cambridge, MA: MIT Press, 167-224.

Boersma, P. & Weenink, D. (2003), *Praat: doing phonetics by computer* (V. 4.0.4), http://www.praat.org/

Brown, C. (2000), The Interrelation between speech perception and phonological acquisition from infant to adult, in *Second Language Acquisition and Linguistic Theory* (J. Archibald, editor), Oxford: Blackwell Publishers, 4-63.

Cantín M. & Ríos A. (1991), Análisis experimental del ritmo de la lengua catalana, *Anuario del Seminario de Filología Vasca Julio de Urquijo / International Journal of Basque Linguistics and Philology* (ASJU), 25, 487-514.

Cutler, A., Mehler, J., Norris, D. & Segui, J. (1986), The syllable's differing role in the segmentation of French and English, *Journal of Memory and Language*, 25, 385-400.

Dahan, D. & Bernard, J.M. (1996), Interspeaker variability in emphatic accent production in French, *Language and Speech*, 39, 341-374.

Dupoux, E., Pallier, C., Sebastián, N. & Mehler, J. (1997), A destressing 'deafness' in French?, *Journal of Memory and Language*, 36, 406-421.

Dupoux, E., Peperkamp, S. & Sebastián, N. (2001), A robust method to study stress 'deafness', *Journal of the Acoustical Society of America*, 110, 1606-1618.

Dupoux, E., Sebastian-Galles, N. Navarete, E. & Peperkamp, S. (2008), Persistent stress 'deafness': the case of French learners of Spanish, *Cognition*, 106, 682-706.

Flege, J. & Hillenbrand, J. (1986), Differential use of temporal cues to the /s/-/z/ contrast by native and non-native speakers of English, *Journal of the Acoustical Society of America*, 79, 508-517.

Flege, J. (1995), Second language speech learning: Theory, findings, and problems, in, *Speech Perception and Linguistic Experience: Issues in Crosslanguage Research* (W. Strange, editor), Baltimore: York Press, 233-273.

Kuhl, P.K. (1991), Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not, *Perception & Psychophysics*, 50, 93-107.

Kuhl, P.K. (2000), A new view of language acquisition, in *Proceedings of the Academy of Sciences of the United States of America*, 97 (22), 11850-11857.

Krashen, S.D. (1987), *Principles and Practice in Second Language Acquisition*, London: Prentice-Hall International.

Llisterri, J., Machuca, M., de la Mota, C., Riera, M. & Ríos, A. (2005), La percepción del acento léxico en español, *Filología y lingüística* 1 (Madrid: CSIC–UNED–U. de Valladolid), 271-297.

Manchón Ruiz, R.M. (1987), Adquisición/aprendizaje de lenguas: el problema terminológico, *Cuadernos de Filología Inglesa*, 3, 37-47.

McAllister, R., Flege, J.E. & Piske, T. (2002), The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian, *Journal of Phonetics*, 30, 229-258.

Mora, E., Courtois, F. & Cavé, C. (1997), Etude comparative de la perception par des sujets francophones et hispanophones de l'accent lexical en espagnol, *Revue Parole*, 1, 75-86.

Muñoz, M., Panissal, N., Billières, M. & Baqué, L. (2008), ¿La metáfora de la criba fonológica se puede aplicar a la percepción del acento léxico español? Estudio experimental con estudiantes francófonos, in *Actas del XXIV Congreso Internacional de la Asociación Española de Lingüística Aplicada*, Almería, Spain, April 3-5.

Otake, T., Hatano, G., Cutler, A. & Mehler, J. (1993), Mora or syllable? Speech segmentation in Japanese, *Journal of Memory and Language*, 32, 258-278.

Peperkamp, S., Dupoux, E. & Sebastián-Gallés, N. (1999), Perception of stress by French, Spanish and bilingual subjects, in *Proceedings of Eurospeech '99.* 6th European Conference on Speech Communication and Technology (vol. 6), Budapest, Hungary, September 5-9, 2683-2686.

Ramus, F., Nespor, M. & Mehler, J. (1999), Correlates of linguistic rhythm in the speech signal, *Cognition*, 73, 265-292.

Rigault A. (1962), Rôle de la fréquence, de l'intensité et de la durée vocalique dans la perception de l'accent en français, in *Proceedings of the 4th International Congress of Phonetic Sciences* (A. Sovijärvi & P. Aalto, editors), The Hague: Mouton, 735-748.

Russo, R. & Barry, W.J. (2008), Isochrony reconsidered. Objectifying relations between rhythm measures and speech tempo, in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, May 6-9, 419-422.

Wang, Y., Spence, M., Jongman, A. & Sereno, J. (1999), Training American listeners to perceive Mandarin tones, *Journal of the Acoustical Society of America*, 106, 3469-3658.

Wang, Q. (2008), L2 stress perception: The reliance on different acoustic cues, in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, May 6-9, 635-638.

PERSISTENZA DELL'ACCENTO STRANIERO. UNO STUDIO PERCETTIVO SULL'ITALIANO L2

Giovanna Marotta [a], Philippe Boula de Mareüil [b]
[a] Dipartimento di Linguistica, Università di Pisa, [b] LIMSI-CNRS, Orsay
gmarotta@ling.unipi.it, mareuil@limsi.fr

1. SOMMARIO

La ricerca fonologica sull'acquisizione di L2 si è finora concentrata sul versante della produzione, trascurando quello della percezione, nonostante sia da tempo nota la rilevanza dei processi percettivi anche nella resa fonetica dei segmenti; in particolare, risulta ancora poco indagata la tematica relativa alla percezione del cosiddetto 'accento straniero'.

All'interno di questa area di ricerca, un problema specifico concerne il peso dei tratti fonetici sulla percezione del *foreign accent*. Presentiamo qui i risultati di un test percettivo in cui alcuni frammenti di parlato italiano prodotto da parlanti con diversa L1 (francese, spagnolo, tedesco, inglese) sono stati valutati da parlanti nativi italiani.

I soggetti sono stati chiamati ad ascoltare gli stimoli acustici naturali, uno per volta, e a giudicare se il parlante era italiano o straniero; se valutato come straniero, i soggetti dovevano indicare la lingua madre del parlante tra le quattro lingue sopra elencate, valutando anche il grado di accento straniero su una scala a tre gradini, che va da 0 (poco accento) a 2 (accento forte). Ogni ascoltatore italiano è stato preliminarmente invitato ad autovalutare sia la sua competenza nelle quattro lingue straniere, che il suo grado di familiarità con l'accento delle stesse lingue.

I risultati mostrano che nella maggioranza dei casi gli ascoltatori sono in grado di percepire la differenza tra parlanti italiani nativi e parlanti non nativi, anche in caso di ottima competenza dell'italiano. Più complesso si è invece rilevato il compito relativo all'identificazione della lingua materna dei parlanti. Soltanto gli stimoli prodotti da parlanti inglesi sono stati identificati con una percentuale di riconoscimento soddisfacente, mentre quelli relativi a parlanti spagnoli presentano i valori di corretto riconoscimento più bassi. Inoltre, gli stimoli prodotti da parlanti tedeschi sono stati spesso confusi con quelli relativi ai parlanti inglesi.

Il grado di successo nel riconoscimento della L1 appare dunque inversamente proporzionale alla vicinanza strutturale e fonologica tra L1 e L2: italiano e spagnolo sono discriminati con difficoltà, mentre il parlato dei tedeschi tende ad essere confuso con quello degli inglesi più che con quello degli spagnoli.

Tuttavia, dai nostri dati non risulta una buona corrispondenza tra l'autovalutazione dell'ascoltatore e la sua *performance* nel test percettivo. In maniera abbastanza prevedibile, soltanto nel caso dell'inglese si osservano valori comparabili tra autovalutazione e percezione, mentre per le altre lingue straniere si rileva una discrasia più o meno marcata tra il supposto livello di familiarità con un accento straniero e la corretta identificazione della lingua straniera nel test sperimentale. In altri termini, la percezione dell'accento straniero può esser indipendente dal corretto riconoscimento della lingua materna parlata da colui che è stato identificato come straniero.

2. INTRODUZIONE

Nonostante da tempo i processi percettivi siano riconosciuti come rilevanti nel processo dell'apprendimento di L2 e quindi nella resa fonetica dei relativi segmenti (cfr. Best, 1995; Flege, 1997 e 2003; Major, 2001), gli studi fonologici sull'acquisizione di lingue seconde si sono finora concentrati quasi esclusivamente sul versante della produzione. All'interno del versante percettivo, il cosiddetto 'accento straniero' rappresenta al momento un tema ancora poco indagato.[1]

Se consideriamo gli studi finora prodotti sull'italiano come L2, possiamo facilmente osservare come la percezione sia terreno rimasto essenzialmente inesplorato. Del resto, la marginalità delle analisi percettive nella letteratura fonetico-fonologica è ben nota, ma non per questo meno deprecabile. Pur tuttavia, un certo cambio di rotta sembra profilarsi all'orizzonte in questi ultimi anni, come risulta dalla compulsazione della letteratura prodotta in tema e riassunta in un nostro recente lavoro (cfr. Marotta, 2008a).

All'interno del vasto campo dedicato all'acquisizione di lingue seconde, un problema specifico concerne il peso dei tratti fonetici, segmentali e soprasegmentali, nella percezione del *foreign accent*. Le domande fondamentali che a nostro avviso dovrebbero essere inserite nell'agenda ideale in merito a questo argomento sono le seguenti:

- i tratti responsabili di 'forestierismo' permangono anche nella produzione di parlanti con ottima competenza di L2?
- in che misura questi tratti dipendono dalle caratteristiche di L1?
- qual è il ruolo dei fattori prosodici nella percezione dell'accento straniero?

Per tentare di rispondere a queste domande, abbiamo programmato una serie di test percettivi, diversi nella composizione degli stimoli e nella loro presentazione, il cui fine sarebbe quello di consentirci di individuare gli elementi che guidano la percezione e di valutare il peso relativo degli elementi segmentali e prosodici nel riconoscimento dell'accento straniero.

Presenteremo qui i risultati di un test percettivo in cui alcuni frammenti di parlato italiano (letto e spontaneo) prodotto da parlanti con diversa L1 (francese, spagnolo, tedesco, inglese) e ottima competenza dell'italiano sono stati valutati da parlanti nativi italiani.

Come vedremo, percepire tratti di 'forestierismo' è relativamente semplice per i nativi, mentre più complesso risulta individuare la lingua straniera, anche in presenza di costante esposizione ad essa.

3. SULLA PERCEZIONE

Produzione fonetica e percezione uditiva non hanno avuto nel corso del tempo destini simili, ma piuttosto diverse fortune. La scarsità di studi dedicati alla percezione interessa sia la fonetica segmentale che quella soprasegmentale. Si osservi a riprova che nel volume *Handbook of Phonetic Sciences* curato da Hardcastle e Laver (1997), sono ben pochi i saggi che si occupano di percezione; in campo prosodico, viene di solito preso in esame soltanto l'accento lessicale (cfr. McQueen & Cutler, 2007) o il ritmo (cfr. Ramus & Mehler, 1999; Ramus *et al*. 1999). Parimenti, si ricordi che il manuale dedicato a *Speech Perception* (Pisoni & Remez, 2005) è del 2005, quindi piuttosto recente, a conferma del tardivo interesse dei linguisti e dei fonetisti in particolare, nei confronti delle tematiche percettive.

[1] Si vedano tuttavia Archibald (1993), Magen (1998), Jilka (2000; 2007).

Le ragioni del ritardo degli studi percettivi, innegabile tanto in Italia quanto nel resto del mondo, sono molteplici. È probabile che una prima motivazione vada ricercata nella differenza sostanziale che sussiste tra *parlare* e *ascoltare*: Albano Leoni (2001) ha giustamente ricordato a questo proposito che *parlare* è un atto esterno e visibile, mentre *ascoltare* è un atto interno e invisibile: si può vedere una persona che parla (anche in condizioni di rumore), mentre non possiamo sapere se una persona sta ascoltando oppure no. La stessa autopercezione è del resto diversa: mentre parliamo, attraverso l'analisi propriocettiva abbiamo coscienza dei nostri organi fonatori (bocca, labbra, lingua), mentre quando ascoltiamo, non possiamo né vedere né percepire i movimenti dell'apparato uditivo.

Non è dunque casuale la circostanza per cui sia la tradizione grammaticale che le prime analisi scientifiche dedicate alle lingue indoeuropee, abbiano dedicato ampio spazio alla fonetica articolatoria, offrendo solo raramente riflessioni e osservazioni sul fronte percettivo.

Per lo studio della percezione linguistica è finora risultata scarsamente rilevante la stessa psicoacustica, disciplina virtualmente mirata alla percezione, segnatamente all'individuazione delle soglie di percettibilità dei parametri fisici (tempo, ampiezza e frequenza; cfr. House, 1990; Moore, 1997): buona parte dei materiali impiegati in questo settore non sono infatti linguistici; ad es., per quanto concerne il parametro della frequenza, spesso si tratta di toni puri. Essendo il nostro ascolto vario e nel contempo altamente specifico, l'orecchio si sintonizza in modo diverso a seconda che si tratti di sequenze di suoni linguistici oppure di rumori o toni puri; di conseguenza, l'analogia potrebbe non funzionare in maniera perfetta, o quanto meno adeguata.

Sul piano teorico, una questione di primaria rilevanza riguarda il tipo di percezione attiva nell'ascolto, sia in generale che nello specifico di catene foniche. Il contrasto classico, ma non di meno essenziale, ruota intorno ai due poli olistico e analitico o lineare. In linea di principio, possiamo dire che il primo tipo di percezione avviene nello spazio, mentre il secondo, nel tempo; ad es. un segnale stradale viene percepito e decodificato mediante la visione in maniera globale, non attraverso la scomposizione dei suoi tratti specifici.[2] Un enunciato oralmente prodotto da un parlante invece si sviluppa linearmente, nel tempo (cfr. tra gli altri Reddy, 1975). Tuttavia, andrà osservato che anche nell'udito si osserva la tendenza a ricomporre l'unità, l'intero inteso come unità di senso. I due tipi di percezione, olistica e analitica interagiscono strettamente, tanto che si può determinare agevolmente il passaggio da un piano percettivo all'altro. Così ad es., si ha passaggio dal lineare all'olistico nei casi di *priming* lessicale; parallelamente, si può passare dall'olistico al lineare, qualora l'occhio e la mente scompongano un segnale stradale nei suoi tratti salienti.

Compiere studi di carattere percettivo pone inoltre una serie di problemi metodologici. Innanzitutto, nei test percettivi, è necessario porre la massima attenzione a COME si formula la richiesta, in modo da evitare la circolarità. In secondo luogo, onde poter controllare che la risposta del soggetto non sia casuale, è opportuno calibrare la durata del test, ed evitare sedute troppo lunghe, che rischierebbero di rendere i risultati non affidabili, in quanto casuali.

Un altro aspetto teorico rilevante concerne il carattere categoriale dei processi percettivi. Che vi sia percezione categoriale nel linguaggio pare incontestabile, ed è stato del resto ampiamente dimostrato a partire dagli studi, per molti versi pionieristici, condotti sul *Voice Onset Time* (cfr. Lieberman *et al*. 1957; 1967; Liberman & Blumstein, 1988). Di recente,

[2] In merito ai processi percettivi, vera e propria galassia di ricerca, si vedano quali primi testi di riferimento Rookes & Willson (2000), Contessi *et al*. (2002).

Harnad (2005), tra gli altri, è tornato sul tema con argomenti che ci paiono convincenti. Ma basterà richiamare alla memoria la nozione classica di fonema per dedurne che la categorizzazione è processo insito nella natura stessa del sistema linguistico.

La questione non è pertanto se nell'ascolto fonetico, cioè linguistico, la percezione possa avere carattere categorico, ma se la percezione sia sempre e solo categoriale. In particolare, in ambito prosodico, esistono categorie prosodiche oppure solo *continua* ?[3]

Consideriamo il caso della lingua italiana: l'accento lessicale in italiano è distintivo, per cui il suo carattere categorico non sembra discutibile. Ma l'intonazione in una lingua non tonale come l'italiano viene parimenti percepita in termini categoriali? A nostro parere, la risposta a questo quesito ha da essere negativa, dal momento che l'unica funzione distintiva che si può riconoscere a livello grammaticale alla melodia nella nostra lingua riguarda l'espressione della domanda polare (cfr. in merito e motivatamente, Marotta, 2002-2003; 2008b).

4. PERCEZIONE: DATI ITALIANI

Gli studi dedicati alla percezione fonetica, sia segmentale che soprasegmentale, in lingue diverse dall'italiano cominciano ad essere copiosi, anche se molti fenomeni non sono stati ancora trattati con sufficiente attenzione.[4]

Se consideriamo la produzione scientifica relativa alla lingua italiana ed alle sue varietà, possiamo osservare che gli studi in questo settore hanno interessato in tre aree di ricerca, con finalità almeno in parte differenti:

- studi sui tratti prosodici che contribuiscono all'identificazione del parlante; cfr. Interlandi (2004), Marotta *et al.* (2004), Marotta & Sardelli (in stampa), Boula de Mareüil *et al*. (2004a, 2004b e 2009), Calamai & Ricci (2005); Gamal (2006 e 2007);
- studi sulla percezione delle categorie prosodiche fonologiche; cfr. Gili Fivela (2004), Savino *et al*. (2006);
- studi sulla percezione segmentale; cfr. Cerrato *et al.* (1994), Albano Leoni *et al*. (1996), Calamai & Ricci (2005), Mori (2007), Celata (2009), Sorianello (2009; in corso di stampa), Avesani *et al*. (2009).

Rappresenta invece ancora un terreno vergine lo studio dell'accento straniero (d'ora in poi AS) da parte di parlanti-ascoltatori italiani. In particolare, non è stato finora indagato a fondo il ruolo dei tratti segmentali e soprasegmentali nella percezione dell'AS, anche se lo stesso uso del sintagma 'accento straniero' lascia intendere che un parlante viene identificato come straniero, cioè non nativo, proprio per il suo accento.[5] Se è vero che gli studi dedicati alla percezione di AS sono scarsi, ancora più scarse sono le indagini sul ruolo svolto dai fattori soprasegmentali nella percezione dell'AS.

---

[3] Per un primo orientamento, si possono vedere i contributi di House (1990), Ladd (1996), Gussenhoven (2002 e 2004), Vaissière (2005).

[4] Per una rassegna, ci permettiamo di rinviare a Marotta (2008a).

[5] L'attributo *straniero* è qui da intendersi nella sua doppia valenza, vale a dire sia in riferimento ad una L2 rispetto ad una L1 che per le varietà regionali di una stessa lingua. Ad esempio, è esperienza comune del parlante italiano la percezione di 'accento regionale' diverso dal proprio, con conseguente riconoscimento dell'area di provenienza dell'interlocutore.

In effetti, i modelli teorici più in voga per lo studio dell'acquisizione di lingue seconde sono concentrati sulla produzione e percezione dei segmenti, anche se sono consapevoli del ruolo della percezione. [6] Eppure, non mancano certo i motivi per cui la prosodia potrebbe essere prioritaria, data la maggiore plasticità della struttura melodica dell'enunciato (cfr. Ladd 1996). D'altra parte, proprio perché le caratteristiche prosodiche della lingua materna sono acquisite molto precocemente dai bambini, assai prima del lessico e della stessa fonologia segmentale, i tratti prosodici corrispondenti si 'fissano' prima, meglio e più rigidamente nella competenza dei parlanti. Ciò spiegherebbe il fatto che anche in soggetti che parlano una lingua straniera a livelli di competenza particolarmente elevati si osservano di frequente alcuni elementi che li caratterizzano comunque come stranieri all'orecchio dei parlanti nativi.

Vi sono quindi fondati motivi per ritenere che la prosodia possa essere una spia potente di AS. La questione essenziale a questo proposito può essere riassunta nei termini seguenti: qual è il ruolo dei tratti prosodici nell'identificazione di un parlante come 'straniero'? Si tratta di un ruolo basico o sussidiario?

A questa domanda abbiamo cercato di rispondere, anche se in via preliminare, con un nostro studio di qualche anno fa (cfr. Boula de Mareüil *et al*., 2004), in cui abbiamo messo a confronto la percezione di frasi italiane e spagnole sia originali che sintetiche, in cui avevamo mescolato la fonetica segmentale originale di L1 con la prosodia di L2. I risultati ottenuti mostravano con chiarezza che la solidarietà tra parametri, segmentali e prosodici, facilita l'identificazione della lingua; questo valeva sia per i soggetti italiani che per quelli spagnoli. Ma nel caso in cui i parametri si incrociavano, cioè, segmenti originali in *Lx* + prosodia sintetica in *Ly*, o viceversa, segmenti sintetici in *Lx* + prosodia originale in *Ly*, si riscontrava la maggiore rilevanza della prosodia rispetto ai segmenti.

Lo studio che qui presentiamo si inserisce lungo la medesima linea di ricerca, ma è centrata sull'AS, dal momento che si focalizza sulla percezione dei parlanti nativi di italiano nei confronti di soggetti che siano in possesso di ottima competenza di italiano come L2.

5. L'ESPERIMENTO PERCETTIVO

Il nostro esperimento percettivo è volto a verificare la percezione dell'accento straniero da parte di parlati nativi italiani nel parlato italiano prodotto da soggetti che risiedono in Italia da lungo tempo. I campioni di parlato che costituiscono la base empirica di ascolto per gli uditori sono tratti da brani di conversazione spontanea e dalla lettura di un articolo di giornale.

5.1 Locutori

Sono stati scelti quattro accenti stranieri piuttosto familiari, ovvero francese, inglese, tedesco e spagnolo; per ognuno di essi sono state selezionate due locutrici di sesso femminile, per un totale di otto parlanti, cui si sono aggiunte due locutrici italiane di controllo, di area toscana nord-occidentale.

I soggetti stranieri che hanno prodotto i materiali oggetto di ascolto risiedono in Italia da molti anni ed in prevalenza svolgono attività di insegnamento della loro madre-lingua presso l'Università di Pisa. Pur essendo tutte in possesso di ottime competenze in lingua

[6] Ciò vale, sia pure in maniera diversa e progressivamente minore, per i modelli *Speech Learning Model* (Flege, 1995, 1997 e 2003; MacKay *et al*., 2001); *Ontogeny and Philogeny Model* (Major 2001) e *Perceptual Assimilation Model* (Best 1995).

italiana, per ogni coppia di parlanti è possibile individuare un certo scarto, nel senso che una delle due (*Loc I*) presenta una produzione leggermente migliore rispetto all'altra (*Loc II*).

5.2 Protocollo di registrazione e preparazione degli stimoli

Le registrazioni sono state effettuate nel Laboratorio di Fonetica e Fonologia del Dipartimento di Linguistica dell'Università di Pisa attraverso microfoni professionali a cravatta SONY e un registratore DAT SONY. La campionatura è stata effettuata a 22.050 Hz con *software Multi-Speech* 3700 (versione 2.5).

Per tutti i soggetti è stato usato lo stesso protocollo sperimentale in modo da avere materiale coerente: per ogni parlante, abbiamo ottenuto in media venti minuti di materiale audio registrato, contenente produzione orale, sia spontanea che letta.

In apertura di seduta di registrazione, è stato chiesto ad ogni soggetto di fare una breve autopresentazione, prima in lingua madre, quindi in italiano; successivamente, è stato suggerito di parlare dell'Italia e degli italiani (pregi e difetti); infine abbiamo domandato alle nostre parlanti come pensavano di passare le proprie vacanze estive. Le risposte a queste domande costituiscono il materiale spontaneo.

Nella seconda parte della seduta di registrazione, abbiamo chiesto alle locutrici di leggere due volte, nel modo più spontaneo possibile, un breve brano estratto da un articolo del quotidiano *Il Corriere della Sera*, afferente ad uno stile formale, con lessico e sintassi di rango elevato, tanto che la lettura del brano si è rivelata un compito piuttosto arduo per i nostri soggetti.

Dopo l'ascolto accurato del materiale registrato, abbiamo provveduto a selezionare per ogni locutore quattro frammenti delle lingue prese in esame (quattro straniere e l'italiano), due relativi al parlato spontaneo e due al parlato letto. Il nostro test si compone pertanto di quaranta stimoli, otto per ogni lingua, di diversa durata (cfr. *ultra*, § 7).

Abbiamo estratto inoltre nove frammenti aggiuntivi per realizzare un esperimento di *training* da sottoporre agli ascoltatori, come fase preparatoria al test vero e proprio.

Per il trattamento e la segmentazione del materiale registrato, con conseguente selezione dei frammenti da sottoporre all'ascolto, ci siamo avvalsi del *software PRAAT* (cfr. www.praat.org).

5.3 Ascoltatori

Il test è stato sottoposto a 127 ascoltatori, prevalentemente studenti dell'Università di Pisa, appartenenti a diversi Corsi di Laurea (*Linguistica, Comunicazione Pubblica, Sociale e d'Impresa, Informatica Umanistica* e *Ingegneria*) di età compresa perlopiù tra i 20 e i 30 anni. Una settantina di soggetti è di area toscana, mentre i restanti partecipanti al test provengono da varie regioni d'Italia. Undici ascoltatori sono lavoratori, la cui l'età è mediamente superiore a quella degli studenti (si tratta in genere di quarantenni, con qualche cinquantenne), tutti di area toscana.

5.4 Protocollo sperimentale del test percettivo

Il protocollo sperimentale seguito nel test prevedeva per ogni ascoltatore la compilazione preliminare di una scheda sociolinguistica, in cui venivano chieste le seguenti informazioni:

- generalità anagrafiche dell'ascoltatore;
- quali fossero le lingue straniere da lui conosciute e da quanto tempo;
- eventuali esperienze di studio o lavoro all'estero;
- il livello di conoscenza che ogni ascoltatore supponeva di possedere nelle lingue straniere francese, inglese, spagnolo e tedesco;
- il grado di confidenza con l'accento relativo alle quattro lingue straniere prima elencate.

Abbiamo ritenuto necessario tenere distinte le due ultime variabili (autovalutazione della conoscenza di L2 e del relativo 'accento'), dal momento che l'identificazione di un accento straniero può prescindere dalla comprensione di una lingua straniera. Ad esempio, nell'Italia contemporanea, molti italiani sono in grado di riconoscere un accento straniero genericamente slavo, magari classificandolo erratamente come russo, pur senza avere la minima competenza, né passiva né attiva, del russo, ma semplicemente perché vi riconoscono alcuni suoni diversi dai propri ed associati a parlanti provenienti da quell'area europea grazie a contatti, anche cursori e temporanei, avuti con loro.

Dopo una breve fase di *training*, ha avuto luogo l'esperimento percettivo vero e proprio con le seguenti consegne:

- indicare la lingua madre del locutore, scegliendola da una lista chiusa comprendente italiano, francese, inglese, spagnolo, tedesco, oppure selezionando 'straniera';
- valutare il grado di accento straniero percepito su una scala a tre valori (nessun accento; accento modesto; accento forte).

Data la brevità degli stimoli (cfr. in media, ogni stimolo durava 4,5 sec.; su questo punto, cfr. *supra* ed *infra*, § 7) è stata data la possibilità di ascoltare ogni frammento anche più di una volta. Dal momento che la percezione di AS è direttamente proporzionale alla lunghezza della sequenza fonica che viene ascoltata dal parlante nativo (cfr. Jilka 2007, *et ultra,* § 7), abbiamo scelto deliberatamente di operare con stimoli uditivi brevi**,** per evitare l'occorrenza di tratti 'forestieri' plurimi e/o rinforzati per iterazione.[7]

[7] Il test è stato reso disponibile *on-line* e gestito elettronicamente sul *web*, mediante l'uso di una piattaforma messa a punto presso il Laboratorio *LIMSI* del *CNRS* di Orsay (France). La registrazione, la segmentazione dei materiali e la preparazione degli stimoli sono state svolte da Susanna Bertucci nell'ambito della sua tesi di Laurea Triennale in *Comunicazione Pubblica, Sociale e d'Impresa* (*P.S.I.*) presso la Facoltà di Lettere dell'Università di Pisa.

6. RISULTATI

In questa sezione del lavoro analizzeremo i risultati emersi dal test percettivo. La *performance* degli ascoltatori sarà valutata in rapporto alle seguenti variabili:

- Tipo di parlato (letto versus spontaneo)
- Grado di accento straniero percepito
- Autovalutazione della conoscenza di L2
- Autovalutazione del riconoscimento dell'AS
- Classe di ascoltatori (lavoratori *versus* studenti)

6.1 Tipo di parlato (letto versus spontaneo)

Esaminiamo in primo luogo i dati relativi all'ascolto degli stimoli di parlato letto.

Analizzando le percentuali riportate nella Tabella 1, possiamo vedere che le parlanti straniere sono state riconosciute dalla maggioranza degli ascoltatori, in media dal 59% per quanto riguarda l'inglese, il francese e il tedesco; le due locutrici italiane sono state identificate in maniera corretta dalla quasi totalità dei partecipanti al test (97%), anche se alcuni ascoltatori le hanno scambiate per tedesche o per francesi. Per quanto riguarda gli stimoli spagnoli invece, la percentuale di ascoltatori che ha riconosciuto la provenienza delle locutrici è poco più del terzo (34%); molti soggetti non hanno notato alcun accento straniero (0,83% contro l'1,44% di accento straniero medio rilevato per inglese, francese e tedesco) e le hanno scambiate per native italiane (24%); inoltre, in una percentuale pari al 22%, le due parlanti spagnole sono state identificate come straniere, ma senza saper indicare con precisione quale fosse la loro lingua madre.

Parlato Letto	*inglese*	*francese*	*tedesco*	*italiano*	*spagnolo*	*straniero*
Inglese	**59**	9	9	9	1	12
Francese	1	**53**	10	19	4	12
Tedesco	15	13	**65**	0	1	7
Italiano	0	1	0	**97**	1	1
Spagnolo	3	11	7	24	**34**	22

Tabella 1: Valori percentuali di risposte corrette nel parlato letto

Analizzando i dati emersi dall'ascolto degli stimoli del parlato spontaneo, possiamo notare sensibili differenze rispetto alla produzione letta: ad eccezione dei risultati prodotti dall'ascolto dei frammenti delle parlanti italiane, per le quali i valori percentuali tra letto e spontaneo sono molto simili (97% e 94%), si osserva un forte aumento di risposte errate nel parlato spontaneo. Nella Tabella 2 si nota come gli stimoli spontanei prodotti da parlanti straniere inglesi sono riconosciuti in una percentuale discreta di occorrenze (43%), mentre quelli prodotti da francesi sono stati individuati con molta difficoltà, dal momento che le risposte corrette sono pari a solo il 22%; inoltre, nella maggior parte delle risposte errate, le locutrici francesi sono state identificate come italiane (45%). Le parlanti spagnole sono le uniche ad essere state riconosciute con più facilità nel parlato spontaneo rispetto a quello letto.

Parlato Spontaneo	*inglese*	*francese*	*Tedesco*	*italiano*	*spagnolo*	*straniero*
inglese	**43**	5	26	9	2	14
francese	3	22	2	**45**	15	12
tedesco	19	25	**36**	8	2	10
italiano	1	1	1	**94**	1	1
spagnolo	2	6	4	26	**53**	9

Tabella 2: Valori percentuali di risposte corrette nel parlato spontaneo

Come è già stato sottolineato, le risposte corrette per quanto riguarda il parlato spontaneo sono inferiori rispetto al letto (tranne che nel caso dello spagnolo). Da un'analisi della varianza di tipo ANOVA, svolta ponendo come variabili dipendenti le risposte date dai soggetti a ciascun stimolo (corretto, con valore 1; sbagliato, con valore 0) e come variabile indipendente il fattore *Stile di parlato* (letto *vs* spontaneo), risulta che lo *Stile* ha un effetto maggiore [$F(1,126) = 149$; $p < 0{,}001$]. Ciò suggerisce che le differenze tra produzione letta e produzione spontanea sono statisticamente significative, indicando che il parlato spontaneo delle nostre locutrici sia di qualità migliore, nel senso di dotato di minore AS rispetto al letto. Torneremo su questo aspetto nel paragrafo conclusivo.

Da un'analisi complessiva dei nostri dati, emerge che le parlanti straniere che sono state riconosciute con una percentuale di successo maggiore sono le locutrici inglesi e tedesche (riconosciute in maniera corretta nel 51% dei casi). Le parlanti francofone, invece, sono state riconosciute in maniera corretta solo nel 37% dei casi; si noti anche che con una percentuale simile (32%), le medesime locutrici sono state riconosciute come native italiane. In nessun'altra lingua straniera si sono registrate due percentuali così prossime tra italiano e lingua straniera. Infine, le spagnole sono state identificate in maniera corretta nel 44% dei casi e per il 25% dei partecipanti al test le loro voci sono state scambiate per quelle di parlanti native italiane. La Tabella 3 sintetizza questi dati.

Totale	*inglese*	*francese*	*tedesco*	*italiano*	*spagnolo*	*straniero*
inglese	**51**	7	18	9	2	13
francese	2	**37**	6	32	10	12
tedesco	17	19	**51**	4	1	8
italiano	1	1	1	**96**	1	1
spagnolo	2	9	5	25	**44**	15

Tabella 3: Valori percentuali di risposte corrette nella produzione totale (parlato letto e spontaneo)

6.2 Grado di accento

Nella Tabella 4, riportiamo i dati relativi alla valutazione del grado di accento straniero percepito dagli ascoltatori, sia in riferimento al parlato letto che a quello spontaneo. Ricordiamo a questo proposito che l'ascoltatore doveva valutare il grado di accento su una scala numerica composta da tre valori: nessun accento = 0; accento modesto = 1; accento forte = 2.

I valori medi ottenuti mostrano che gli ascoltatori hanno percepito un livello di accento straniero maggiore per il tedesco (1,71 e 1,52) e per l'inglese (1,42 e 1,38), tanto nel parlato letto che nello spontaneo.[8]

Grado di accento	Parlato letto	Parlato spontaneo
inglese	**1,42**	**1,38**
francese	1,20	0,93
tedesco	**1,71**	**1,52**
spagnolo	0,83	0,93

Tabella 4: Grado di accento medio percepito negli stimoli per ogni lingua (parlato letto e spontaneo); scala 0-2

Come si ricorderà (cfr. § 5.1), pur essendo le nostre locutrici straniere in possesso di ottime competenze in lingua italiana, per ogni coppia di parlanti di L1 diversa dall'italiano è possibile individuare un certo scarto di competenza e, soprattutto, di 'accento', nel senso che una delle due (di seguito indicata come *Loc I*) presenta una produzione migliore rispetto all'altra (denominata *Loc II*).

La nostra intuizione è confermata dai dati percettivi, dal momento che, se scorporiamo i dati della Tabella 4, mantenendo distinte le valutazioni del grado di AS di *Loc I* da quelle di *Loc II*, possiamo facilmente rilevare una diversa attribuzione di grado di AS. Nella Tabella 5 si nota infatti come i valori relativi alla prima locutrice siano sempre inferiori a quelli della seconda locutrice, sia nel parlato letto che in quello spontaneo. Questi risultati dimostrano pertanto che i parlanti nativi di italiano sono in grado non solo di riconoscere un parlante come straniero, ma anche di percepire sottili gradi di AS.

Grado di accento	Parlato letto		Parlato spontaneo	
	Loc 1	Loc II	Loc 1	Loc II
inglese	0,97	1,89	1,11	1,58
francese	0,75	1,65	0,64	0,70
tedesco	1,55	1,82	1,16	1,52
spagnolo	0,77	0,92	0,91	0,99

Tabella 5: Grado di accento medio percepito negli stimoli delle due diversi locutrici per ogni lingua (parlato letto e spontaneo); scala 0-2

[8] I numeri che compaiono nella Tabella 4 come pure nella Tabella 5 riportano il valore medio rispetto alla valutazione espressa dagli ascoltatori su entrambi gli stimoli, a parità di lingua.

6.3 Autovalutazione dei partecipanti al test

Ricordiamo che prima di iniziare il test percettivo avevamo chiesto ai partecipanti di autovalutare la propria conoscenza delle lingue straniere oggetto di studio e la propria capacità di riconoscere gli accenti stranieri. Nella Tabella 6, presentiamo i valori dell'autovalutazione della conoscenza della lingua e dell'accento in rapporto alle risposte corrette date dagli ascoltatori.

Conoscenza	*Lingua Straniera*	*Accento Straniero*	*Risposte corrette*
inglese	96 %	94 %	51 %
francese	51 %	94 %	37 %
tedesco	29 %	88 %	51 %
spagnolo	42 %	96 %	44 %

Tabella 6: Valori percentuali di autovalutazione della conoscenza della lingua straniera e dell'accento straniero in rapporto alle risposte di corretta identificazione

Consideriamo in primo luogo i dati in riferimento al grado di conoscenza dell'inglese: il 96% degli ascoltatori ha affermato di possedere un livello di conoscenza medio-buono di questa lingua e il 94% ha dichiarato di essere in grado di riconoscere l'accento inglese; ciò nonostante, le risposte corrette fornite dagli ascoltatori sono poco più della metà del totale (51%). Non c'è quindi un rapporto direttamente proporzionale tra autovalutazione delle proprie conoscenze di L2 e corretta identificazione di AS. Lo stesso dato emerge in maniera ancora più evidente e speculare nel caso del tedesco: soltanto il 29% dei soggetti ha affermato di avere una discreta conoscenza di questa lingua straniera, eppure le risposte corrette sono il 51%, vale a dire la stessa percentuale dell'inglese. Più sfumati i risultati per le altre due lingue straniere. Nel caso dello spagnolo, il 42% degli ascoltatori ha affermato di possedere un livello medio-alto di competenza linguistica, mentre la quasi totalità dei soggetti (96%) ha dichiarato di essere capace di riconoscere l'accento spagnolo; tuttavia, le risposte corrette sono solo il 44%. Parimenti contraddittori i dati relativi al francese: ad una percentuale pari al 51% ed al 94% di soggetti che affermano di possedere una conoscenza buona della lingua e dell'accento corrispondente, corrisponde soltanto il 37% di risposte esatte.

Se ora analizziamo i dati relativi all'autovalutazione degli ascoltatori in rapporto alla percezione dell'accento straniero (cfr. Tabella 6), osserviamo valori percentuali molto alti, prossimi o superiori al 90%, il che indica che quasi tutti gli ascoltatori hanno affermato di poter riconoscere facilmente i diversi accenti stranieri. Tuttavia, vista la *performance* ottenuta dai medesimi soggetti, risulta chiaro che i partecipanti al test hanno sopravvalutato non poco la propria capacità di riconoscere un accento straniero; d'altra parte, andrà osservato che molti dei soggetti hanno affermato di aver trovato il compito di riconoscimento abbastanza difficile a causa della brevità degli stimoli.

Al fine di analizzare il contributo dell'autovalutazione del livello di competenza in L2 e della familiarità con l'accento corrispondente, abbiamo contato per ciascuna L2 il numero di risposte corrette ottenute per ciascuno dei 127 soggetti che hanno preso parte al nostro test

percettivo, ottenendo così quattro vettori di dimensione 127, rispettivamente per le lingue inglese, francese, spagnolo e tedesco. Abbiamo quindi calcolato le correlazioni di ciascuno di questi vettori con il vettore dei livelli di competenza dei soggetti in L2 da una parte e dall'altra con il vettore della conoscenza dell'accento corrispondente in italiano. Tutte queste correlazioni sono positive, ma piuttosto deboli: 0,2 per le conoscenze in inglese e in francese; 0,3 per le conoscenze in spagnolo; 0,4 le conoscenze in tedesco. Le correlazioni con i valori di autovalutazione della conoscenza dell'accento straniero sono ancora più deboli, pari a 0,1 per ciascuna lingua, il che indica la mancanza di correlazione tra queste variabili, che risultano statisticamente irrilevanti. L'analisi statistica testé compiuta conferma dunque la discrepanza tra la conoscenza di L2 (anche di tipo diverso) e la capacità dichiarata di riconoscere gli accenti stranieri corrispondenti; tale discrepanza è coerente con la sopravvalutazione della propria competenza già notata per i nostri ascoltatori.

6.4 I diversi gruppi dei partecipanti al test

Passiamo ora a considerare la variabile relativa al gruppo di appartenenza degli ascoltatori. Come si ricorderà, la maggior parte dei nostri ascoltatori sono studenti universitari, mentre soltanto undici di loro sono lavoratori con un'età compresa tra i 22 e i 59 anni e con diversi gradi di istruzione.

Prendiamo in considerazione i dati in relazione agli stimoli letti e spontanei, con esclusione dei dati relativi alla lingua italiana, che, come abbiamo già avuto modo di osservare, ottengono percentuali di riconoscimento sempre molto alte, pari o superiori al 90%. Come si evince dalla Tabella 7, il gruppo che ha totalizzato la percentuale maggiore di risposte corrette è quello degli studenti di *Informatica umanistica* (valore medio di risposte corrette 51%). Il valore ottenuto dagli studenti di *Linguistica* (46%) risulta basso rispetto alle aspettative per almeno tre ragioni: prima di tutto, in questo gruppo gli ascoltatori hanno affermato di avere ottime conoscenze della lingua (e questo dato conferma il *mismatch* tra autovalutazione e *performance*; cfr. § precedente), in secondo luogo, molti di questi studenti hanno compiuto esperienze di studio all'estero di almeno sei mesi; infine, ed è il dato più pesante, si tratta di studenti di Laurea Magistrale, quindi con un numero maggiore di anni di studio nelle lingue straniere rispetto agli studenti del triennio. Gli studenti di *Comunicazione P.S.I.* presentano un valore percentuale medio nel riconoscimento degli accenti stranieri presentati simile a quello degli studenti di *Linguistica*, per la precisione, 45%.

Produzione totale	*Lavoratori*	*Stud. Inf.*	*Stud. Ling.*	*Stud. Com.*
inglese	42 %	59 %	35 %	52 %
francese	44 %	39 %	44 %	26 %
tedesco	67 %	61 %	52 %	61 %
italiano	95 %	90 %	99 %	98 %
spagnolo	35 %	47 %	53 %	42 %
valore medio	57 %	59 %	57 %	56 %

Tabella 7:Valori percentuali di risposte corrette dei gruppi di partecipanti al test scorporati per lingua straniera

Un altro dato sorprendente, almeno di primo acchito, riguarda il gruppo comprendente i lavoratori, che ha totalizzato un valore percentuale medio (57%) di risposte corrette del tutto comparabile a quello degli studenti. In teoria, sarebbe stato più probabile aspettarsi risultati decisamente migliori da parte degli studenti, dal momento che i lavoratori che hanno preso parte al nostro test hanno affermato di aver avuto in passato scarsa esperienza con le lingue e attualmente svolgono mansioni nelle quali è presente, e solo saltuariamente, un utilizzo elementare dell'inglese.

Inoltre, nel considerare il valore percentuale di corretta identificazione della lingua straniera in rapporto ai diversi gruppi di ascoltatori, si osservano differenze sensibili da lingua a lingua. In particolare, in riferimento all'inglese, i lavoratori hanno dato nel 42% la risposta corretta, gli studenti di *Informatica* nel 59% dei casi, i comunicatori nel 52% e gli studenti di *Linguistica* appena nel 35%; per quanto riguarda i lavoratori, il dato è in linea con le altre lingue.

Resta da sottolineare che le percentuali più alte di corretta identificazione della lingua straniera sono relative al tedesco, per tutti e quattro i gruppi di ascoltatori presi in esame: nonostante la manifesta e spesso dichiarata incompetenza in questa lingua, gli ascoltatori italiani che hanno preso parte al nostro test, indipendentemente dalla loro provenienza geografica, dal loro genere e dal loro status di studenti o meno, sono stati capaci di riconoscere la lingua materna tedesca delle locutrici nella maggior parte degli stimoli impiegati in questo esperimento percettivo; le percentuali di riconoscimento oscillano infatti per il tedesco tra il 52% e il 67%.

Come già nella sezione 6.1, abbiamo svolto un'analisi statistica ANOVA sulle risposte ottenute, attribuendo il valore 1 alle risposte corrette e 0 alle risposte sbagliate, e considerando come fattori lo stile (letto o spontaneo) e la classe di appartenenza degli ascoltatori; per rendere equilibrato il test, abbiamo considerato 11 soggetti per ciascuna classe. I risultati ottenuti indicano che soltanto il fattore stile è significativo ($F(1,40) = 39.33$; $p > 0.001$), mentre non lo è il fattore classe di appartenenza.

Nel complesso, i dati raccolti mostrano pertanto una sostanziale uniformità di comportamento tra studenti e lavoratori; e tra gli studenti, pari uniformità, dal momento che i valori percentuali di corretto riconoscimento dell'accento straniero sono analoghi e comparabili; soprattutto, non sono statisticamente significativi, almeno per il campione qui considerato: il tasso di successo è tendenzialmente lo stesso, indipendentemente dal corso di laurea seguito dallo studente.

7. DISCUSSIONE

Il nostro esperimento ha dimostrato che la sensibilità percettiva dei parlanti nativi è assai fine: riconoscere una produzione non nativa è compito relativamente semplice, tanto che basta un solo indice fonetico, segmentale o prosodico, tipico di L1 e persistente in L2, per trasmettere la percezione di AS. La percezione di non natività risulta dunque dimostrata dai dati ottenuti in questa sede: gli ascoltatori italiani nativi sono capaci di riconoscere un accento straniero in quasi tutti gli stimoli presentati, nonostante siano stati prodotti da parlanti con un'ottima competenza dell'italiano (cfr. Vaissière 2005; Trouvain & Gut 2007). Del resto, è noto da tempo che la L1

costituisce una sorta di 'filtro' per l'acquisizione di L2, per cui tracce di L1, più o meno consistenti e numerose, si mantengono nella produzione di L2. [9]

Assai più difficile si è invece rilevato il compito di identificazione della lingua madre straniera, dal momento che i nostri soggetti non sempre sono stati in grado di individuare correttamente la provenienza dei parlanti percepiti come non nativi. Soltanto gli stimoli inglesi presentano percentuali di riconoscimento discrete, in media intorno al 50%, mentre i valori più bassi si riscontrano in riferimento agli stimoli prodotti da soggetti di madrelingua spagnola, che abbastanza di frequente vengono identificati come italiani. In parallelo, gli stimoli dei parlanti tedeschi sono spesso confusi con quelli inglesi, mentre quelli francesi con quelli italiani.

Dai risultati raccolti in questo esperimento percettivo, abbiamo avuto modo di vedere come, contrariamente a quanto avremmo potuto attenderci, la *performance* dei nostri ascoltatori sia stata migliore nel caso del parlato letto rispetto a quello spontaneo (cfr. § 6.1 e Tabelle 1 e 2). Il modello OPM di Major prevede che in rapporto alla variazione stilistica, le tre componenti di *Interlanguage*, cioè L1, L2 e U,[10] varino in modo direttamente proporzionale al grado di formalità, per cui quanto più lo stile d'eloquio diventa formale, tanto più cresce la competenza in L2 e diminuisce il peso di L1; quanto alla componente Universale, il modello prevede che sia più attiva negli stadi iniziali e in stili informali, mentre dovrebbe essere minore in stadi avanzati di competenza in L2 e negli stili formali.

In particolare, riguardo alla pronuncia, il modello di Major (2001: 93-94) postula che l'influenza del *transfer* sia minore in situazioni formali e che l'accuratezza fonetica cresca proporzionalmente al grado di formalità dello stile. Naturalmente, in questo schema generale entrano anche fattori extralinguistici, come l'ansia che si può scatenare in una situazione formale, oppure il livello di dimestichezza, maggiore o minore, con un certo stile.

Se equipariamo la produzione letta con lo stile formale e la produzione spontanea con quello informale, possiamo forse individuare una causa del mancato effetto della differenza di stile proprio nel fatto che le nostre parlanti hanno maggiore dimestichezza con lo stile formale (non dimentichiamo che si tratta di docenti di Lingua), tanto nella lingua parlata che in quella scritta.

In realtà, riteniamo che nel nostro caso il parlato letto non sia del tutto equiparabile con lo stile formale e che il parlato spontaneo non sia equiparabile con lo stile informale. La lettura del brano prescelto era infatti piuttosto difficile, per cui è probabile che l'attenzione dei soggetti si sia concentrata soprattutto sul cogliere il significato di quanto stavano leggendo, tralasciando i dettagli fonetici della propria produzione in lingua italiana. D'altro lato, la conversazione spontanea delle nostre parlanti si è svolta in ambiente formale, e con persona a loro nota solo superficialmente, per cui il loro eloquio si è mantenuto su un registro stilistico elevato, non propriamente informale.

Ciò detto, riteniamo che una possibile causa del diverso comportamento rilevato nell'ascolto del parlato letto rispetto a quello classificato in questa sede come spontaneo sia da ricercarsi altrove, vale a dire nella durata dei frammenti di parlato proposti. La durata media degli stimoli uditivi è infatti pari a 6 sec. nel caso del parlato letto, ma solo a 3,5 sec. nel caso del parlato spontaneo. La minore durata degli stimoli relativi al parlato spontaneo potrebbe pertanto

[9] Cfr. già Trubeckoj (1958); più recentemente, a parte i sopra citati modelli di acquisizione di L2 (SLM, OPM e PAM), si vedano i lavori di Dupoux *et al*. (1997) e di Dupoux & Peperkamp (2002).

[10] Nel modello OPM, U sta per *Universal Grammar*, vale a dire princìpi generali e restrizioni presenti in tutte le lingue.

spiegare perché i soggetti abbiano avuto mediamente più difficoltà nel compito di identificazione dell'AS nel caso del parlato spontaneo rispetto a al parlato letto.

Verrebbe quindi confermata l'ipotesi avanzata da Jilka (2007): non solo la percezione di AS, ma anche la capacità di identificazione di L1 sono direttamente proporzionali alla lunghezza della sequenza fonica che viene ascoltata dal parlante nativo. Del resto, gli stessi soggetti hanno talvolta osservato al termine del test che i frammenti sonori proposti erano troppo brevi per consentire loro di svolgere in maniera adeguata il compito richiesto.

Nell'analisi dei risultati ottenuti, abbiamo altresì visto che l'autovalutazione della propria capacità di riconoscere l'accento relativo alle quattro lingue straniere in esame non ha sempre trovato un buon grado di corrispondenza con i risultati ottenuti dai singoli soggetti nel compito di riconoscimento del parlante straniero. L'accento inglese è quello che risulta approssimare di più gli indici di autovalutazione e quelli di percezione,[11] mentre per le altre lingue si assiste ad una più o meno marcata discrasia tra il presunto livello di confidenza con un accento straniero e l'effettivo riconoscimento di quell'accento nel test sperimentale. In altri termini, la percezione di AS prescinde dall'identificazione della lingua parlata dal 'forestiero', come del resto emerge già da studi pregressi (cfr. Boula de Mareüil *et al.* 2004a; 2004b; 2009). Soltanto dopo aver percepito l'AS, cioè un accento diverso dal proprio, il parlante nativo formula le sue ipotesi sulla provenienza linguistica di colui che sta ascoltando, ed è forse in questa seconda fase che emerge il peso della conoscenza della lingua straniera e della familiarità con l'AS.

Sembra pertanto che la competenza dei nativi su una L2 non sia di per sé un fattore in grado di facilitare gli ascoltatori nell'individuazione dell'AS e che il riconoscimento di un AS possa dipendere non tanto dal bagaglio di conoscenze di quella lingua, lessicali e grammaticali, ma dalla capacità di riconoscere i suoi elementi fonetici caratteristici, sia segmentali che prosodici. Tale abilità è in parte dipendente dalla familiarità con la lingua straniera, ad esempio dall'ascolto diretto di parlanti nativi, ma è anche legata alla sensibilità dell'individuo.

Non è ancora chiaro se siano i tratti segmentali o quelli prosodici a giocare un ruolo maggiore nell'identificazione dell'accento straniero; d'altra parte, è anche possibile che non ci sia una risposta assoluta e che la preferenza di uno dei due elementi sia legata a fattori soggettivi. Gli studi condotti da Boula de Mareüil *et al.* (2004a; 2004b; 2009) mostrano in effetti risultati contrastanti: per certi versi, sembrano suggerire che i tratti segmentali giochino il ruolo più importante nella percezione di AS; per altri, invece, parrebbe che la prosodia costituisca l'elemento principale. Studi ulteriori, mirati in questo senso, potranno gettare nuova luce su questo aspetto.

Infine, dai dati raccolti si evince che una maggiore similarità tra L1 e L2 non facilita il compito di riconoscimento di AS. Al contrario, più le lingue sono simili tra di loro, più è difficile discriminare l'AS. Nel nostro campione, ricordiamo che gli stimoli prodotti dalle parlanti spagnole sono risultati i più difficili da identificare ; parimenti, gli stimoli tedeschi sono stati spesso confusi con quelli inglesi. Il grado di successo nel riconoscimento della L1 è dunque inversamente proporzionale alla vicinanza strutturale e fonologica tra L1 e L2: italiano e spagnolo sono discriminati con difficoltà, mentre il parlato dei tedeschi tende ad essere confuso con quello degli inglesi più che con quello degli spagnoli.

[11] E certo questo dato non ci stupisce, visto che l'inglese non solo è la lingua con cui lo studente medio italiano si confronta nel suo pluriennale percorso scolastico, ma anche è la lingua straniera cui viene costantemente esposto attraverso i mass-media.

Lo studio percettivo qui presentato andrà interpretato come una tappa nel percorso di ricerca, lungo e complesso, che dovrebbe consentirci di rispondere alle domande formulate all'inizio di questo lavoro. In particolare, dato l'assetto metodologico di questo esperimento, non è possibile avanzare ipotesi documentate e articolate in merito alla valutazione del peso relativo degli elementi segmentali e prosodici nel riconoscimento dell'AS.

Pur tuttavia, i risultati raccolti ci consentono di giungere alle seguenti e parziali conclusioni, distinte in rapporto al piano di analisi:

- in produzione di L2, è assai difficile perdere completamente l'*imprinting* fonetico-prosodico dato dalla prima lingua;
- sul piano percettivo, la sensibilità dei parlanti nativi nei confronti della propria lingua è talmente fine che è sufficiente la persistenza anche di pochi tratti – segmentali e/o soprasegmentali – percepiti come non appartenenti alla propria lingua per trasmettere il percetto di 'accento straniero';
- a livello fonologico, il grado di successo nel riconoscimento di una lingua straniera è inversamente proporzionale alla vicinanza strutturale e fonologica tra L1 e L2.

Ulteriori studi, condotti su questa medesima linea di ricerca o su vie parallele, potranno in futuro confermare o meno queste conclusioni, nella misura in cui potranno gettare nuova luce su un argomento tanto complesso quanto intrigante, quale è quello dei rapporti tra struttura segmentale e struttura soprasegmentale nella percezione del parlato.

8. BIBLIOGRAFIA

Albano Leoni, F. (2001), Il ruolo dell'udito nella comunicazione linguistica. Il caso della prosodia, *Rivista di Linguistica*, 13, 45-68.

Albano Leoni F., Cutugno F. & Savy R. (1996), The vowel system of Italian connected speech, in P. Branderud & K. Elenius (eds.), *Proceedings of the 13th International Congress of Phonetic Sciences*, Vol. IV, Stockholm: Almqvist & Wiksell, 396-399.

Archibald, J. (1993), *Language learnability and L2 phonology: the acquisition of metrical parameters*, Amsterdam: Kluwer.

Avesani, C., Vayra M., Best, C. & Bohn O.-S. (2009), Fonologia e acquisizione. In che modo l'esperienza della lingua materna plasma la percezione dei suoni del linguaggio?, in *Processi fonetici e categorie fonologiche nell'acquisizione dell'italiano* (L. Costamagna & G. Marotta, editors), Pisa: Pacini, 15-41.

Bernini, G. (1988), Questioni di fonologia nell'italiano lingua seconda, in *L'italiano tra le altre lingue: strategie di acquisizione* (A. Giacalone Ramat, editor), Bologna: il Mulino, 77-90.

Best, C.T. (1995), A direct realistic view of cross-language speech perception, in *Speech perception and linguistic experience* (W. Strange, editor) Baltimore (MD): York Press, 171-206.

Boula de Mareüil, P., Marotta, G. & Adda-Decker, M. (2004a), *Contribution of prosody to the perception of Spanish/Italian accents,* in *Speech Prosody 2004* (B. Bel & I. Marlien, editors), Nara (Giappone), 681-684.

Boula de Mareüil, P., Brahimi, B. & Gendrot, C. (2004b), Role of segmental and supra-segmental cues in the perception of Maghrebian-accented French, in *Proceedings of the 8th International Conference on Spoken Language Processing*, Jeju, 341–344.

Boula de Mareüil, P., Vieru-Dimilescu, B, Woehrling, C. & Adda-Decker, M. (2009), Accents étrangers et régionaux en français. Caractérisation et identification, *Traitement Automatique des Langues* 49/3, in stampa.

Calamai, S. & Ricci, I. (2005), Sulla percezione dei confini vocalici in Toscana: primi risultati, in *Misura dei parametri. Aspetti tecnologici ed implicazioni nei modelli linguistici*, Atti del 1° Convegno Nazionale dell'Associazione Italiana si Scienze della Voce, Padova, 2-4 dicembre 2004 (P. Cosi, editor), Torriana (RN): EDK Editore.

Celata, C. (2009), I contrasti allofonici nella percezione nativa e non-nativa, in *Processi fonetici e categorie fonologiche nell'acquisizione dell'italiano* (L. Costamagna & G. Marotta, editors), 178-221.

Cerrato L., Cutugno F., Frattini, G. & Savy, R. (1994), Un'indagine sulla definizione del confine percettivo tra foni vocalici, in *Atti del 22° Convegno Nazionale dell'AIA*, Lecce, 437-442.

Contessi, R., Mazzeo, M. & Russo, T. (editors) (2002), *Linguaggio e percezione. Le basi sensoriali della comunicazione linguistica*, Roma: Carocci.

Dupoux, E., Pallier, C., Sebastian-Gallés, N. & Mehler, J. (1997), A destressing 'deafness' in French?, *Journal of Memory and Language*, 36, 406-421.

Dupoux, E. & Peperkamp, S. (2002), Fossil markers on language development: phonological deafness in adult speech processing, in *Phonetics, Phonology, and Cognition* (B. Laks & J. Durand, editors), Oxford: OUP, 168-190.

Flege, J.E. (1995), Second language speech learning: Theory, findings, and problems, in *Speech perception and linguistic experience: theoretical and methodological issues* (W. Strange, editor), Timonium, MD: York Press, 229-273.

Flege, J.E. (1997), The Role of Phonetic Category Formation in Second-Language Speech Learning, in *New Sounds 97*, Proceedings of the Third International Symposium on the Acquisition of Second Language Speech, University of Klagenfurt, 8-11 September 1997, 79-88.

Flege, J.E. (2003), Assessing constraints on second-language segmental production and perception, in *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities* (A. Meyer & N. Schiller, editors), Berlin: Mouton de Gruyter, 319-355.

Gamal, D. (2006), La prosodia direttiva in italiano L2. Studio pilota, in *Analisi prosodica. Teorie, modelli e sistemi di annotazione* (R. Savy & C. Crocco, editors), Atti del 2° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Salerno, 30 novembre-2 dicembre 2005, Torriana (RN): EDK Editore.

Gamal, D. (2007), Sul ritmo in italiano L2, in *Scienze vocali e del linguaggio. Metodologie di valutazione e risorse linguistiche* (V. Giordani, V. Bruseghini & P. Cosi, editors), Atti del 3° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, 29 novembre-1°dicembre 2006, Povo (Trento), Torriana (RN): EDK editore, 101-118.

Gili Fivela, B. (2004), La percezione degli accenti: il ruolo dell'allineamento e dello *scaling* dei bersagli tonali, in *Misura dei parametri. Aspetti tecnologici ed implicazioni nei modelli linguistici* (P. Cosi, editor), Atti del 1° Convegno Nazionale dell'Associazione Italiana si Scienze della Voce, Padova, 2-4 dicembre 2004, Torriana (RN): EDK Editore

Gussenhoven, C. (2002), Intonation and Interpretation: Phonetics and Phonology, in *Speech Prosody 2002* (B. Bel & I. Marlien, editors), Université de Provence, Aix-en-Provence, 47-57.

Gussenhoven, C. (2004), *The phonology of tone and intonation*, Cambridge: Cambridge University Press.

Hardcastle, W.J. & Lave J., (editors) (1987)*, The Handbook of Phonetics*, Cambridge: Blackwell.

Harnad S. (2005), To Cognize is to Categorize: Cognition is Categorization, in *Handbook of Categorization in Cognitive Science* (Cohen H. & C. Lefebvre, editors), Amsterdam: Elsevier.

House, D. (1990), *Tonal Perception in Speech*, Lund: Lund University Press.

Interlandi, G. (2004), *L'intonazione delle interrogative polari nell'italiano parlato a Torino: tra varietà regionale e nuova koiné*, Tesi di Dottorato in Linguistica, Università di Pavia.

Jilka, M. (2000), *The contribution of intonation to the perception of foreign accent. Identifying intonational deviations by means of F0 generation and resynthesis*, Ph. D. thesis, University of Stuttgart.

Jilka, M. (2007), Different manifestations and perceptions of foreign accent in intonation, in *Non-Native Prosody. Phonetic description and teaching practice* (J. Trouvain & U. Gut, editors), Berlin – New York: Mouton de Gruyter, 77-96.

Ladd, D.R. (1996), *Intonational phonology*, Cambridge: Cambridge University Press.

Liberman, A.M., Harris, K.S., Hoffman, H.S. & Griffith, B.C. (1957), The discrimination of speech sounds within and across phoneme boundaries, *Journal of Experimental Psychology*, 54, 358-368.

Liberman, A.M., F.S. Cooper, D.S Shankweiler & Studdert-Kennedy, M. (1967), Perception of the Speech Code, *Psychological Review*, 74, 431-461.

Liberman A.M. & Blumstein, S.E. (1988), *Speech Physiology, Speech Perception, and Acoustic Phonetics*, Cambridge (UK): Cambridge University Press.

MacKay, I.R.A, Flege, J.E., Piske, T. & Schirru, C. (2001), Category restructuring during second-language (L2) speech acquisition, *Journal of the Acoustical Society of America*, 110, 516-528.

Magen, H.S. (1998), The perception of foreign-accented speech, *Journal of Phonetics*, 26, 381-400.

Major, R.C. (2001), *Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology*, Mahwah-London: Erlbaum Ass.

Marotta, G. (2002-2003), L'illusione prosodica, *Studi in memoria di Tristano Bolelli, Studi e Saggi Linguistici* XL-XLI, 237-258.

Marotta, G. (2008a), Sulla percezione dell'accento straniero, in *Diachronica et synchronica. Studi in onore di Anna Giacalone Ramat* (R. Lazzeroni, E. Banfi, G. Bernini, M. Chini, G. Marotta, editors), Pisa: ETS, 327-347.

Marotta, G. (2008b), Phonology or non phonology? That is the question (in intonation), *Estudios de Fonetica Expérimental* XVII, 177-206.

Marotta, G., Calamai, S. & Sardelli, E. (2004), Non di sola lunghezza. La modulazione di f0 come indice sociofonetico, in *Costituzione, gestione e restauro di corpora vocali* (A. De Dominicis, L. Mori & M. Stefani, editors), Atti delle XIV Giornate del Gruppo di Fonetica Sperimentale, Roma: Esagrafica, 210-215.

Marotta, G. & Sardelli, E. (in stampa), Prosodiatopia: parametri prosodici per un modello di riconoscimento diatopico, in *Atti del Congresso della Società di Linguistica Italiana* (G. Ferrari & M. Mosca, editors), Vercelli, settembre 2005, Roma: Bulzoni.

McQueen, J. & Cutler, A. (1997), *Cognitive processes in speech perception*, in *The Handbook of Phonetic Sciences* (W.J. Hardcastle & J. Laver, editors), Oxford (UK): Blackwell, 566-585.

Moore, B.C.J. (1997), Aspects of auditory processing related to speech perception, in *The Handbook of Phonetic Sciences* (W.J. Hardcastle & J. Laver, editors), Oxford (UK): Blackwell, 619-639.

Mori, L. (2007), *Fonetica dell'italiano L2*, Roma: Carocci.

Pisoni, D.B. & Remez, R.E. (editors) (2005), *The Handbook of Speech Perception*, Cambridge (UK): Blackwell.

Ramus, F. & Mehler, J. (1999), Language identification with suprasegmentals cues: A study based on speech resynthesis, *Journal of the Acoustical Society of America*, 105, 512-521.

Ramus, F., Nespor,M. & Mehler, J. (1999), Correlates of linguistic rhythm in the speech signal, *Cognition*, 73, 265-292.

Reddy, R.D. (1975), *Speech Recognition*, Berlin: Academic Press.

Rookes, P. & Willson, J. (2000), *La percezione*, Bologna: il Mulino.

Savino, M., Grice, M., Gili Fivela, B. & Marotta, G. (2006), Intonational cues to discourse structure in Bari and Pisa Italian: Perceptual evidence, in *Proceedings of Speech Prosody 2006* (B. Bel & I. Marlien, editors), Dresden 2-5 May 2006, 114-117.

Sorianello, P. (2009), Sulla geminazione in italiano L2, in *Processi fonetici e categorie fonologiche nell'acquisizione dell'italiano* (L. Costamagna & G. Marotta, editors), Pisa: Pacini, 121-146.

Sorianello P. (in stampa), Sull'acquisizione del tratto di lunghezza consonantica nell'italiano L2, in *La Fonetica Sperimentale: Metodo e Applicazioni* (L. Romito, V. Galatà & R. Lio, editors), Atti del 4° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Arcavacata di Rende (CS), 3-5 dicembre 2007, Torriana: EDK Editore.

Trouvain, J. & Gut, U. (editors) (2007), *Non-Native Prosody. Phonetic description and teaching practice*, Berlin – New York: Mouton de Gruyter.

Trubeckoj, N.S. (1958), *Grundzüge der Phonologie*, Göttingen: Vandenhoeck & Ruprecht; trad. it. [1971], *Fondamenti di fonologia*, Torino: Einaudi.

Vaissière, J. (2005), *Perception of intonation*, in *The Handbook of Speech Perception* (D.B. Pisoni & R.E. Remez, editors), Malden (MA)-Oxford (UK): Blackwell, 236-263.

PERCEZIONE E PRODUZIONE DEI FONEMI DELL'INGLESE AMERICANO IN PARLANTI CON UN SISTEMA PENTAVOCALICO

Bianca Sisinni, Mirko Grimaldi [1]
Centro di Ricerca Interdisciplinare sul Linguaggio (CRIL), Università del Salento
bianca.sisinni@unisalento.it, mirko.grimaldi@unisalento.it

1. SOMMARIO

In questo lavoro sono stati analizzati i processi di produzione e percezione durante l'acquisizione dei fonemi vocalici dell'Inglese Americano (AE) in un gruppo di studenti universitari della Facoltà di Lingue dell'Università del Salento parlanti nativi dell'Italiano Salentino (IS) con un sistema a 5 vocali e tre gradi di apertura. Poiché in letteratura è ancora dibattuta la questione se lo sviluppo percettivo della L2 preceda lo sviluppo della produzione, o se una idonea percezione non sia condizione necessaria per una corretta produzione, abbiamo anche cercato di capire qual è il rapporto che intercorre fra il livello percettivo e quello articolatorio (cfr. revisioni in Listerri, 1995; Leather, 1999; Escudero, 2005; Hansen Edwards & Zampini, 2008). Infine, abbiamo verificato se il framework di uno dei modelli percettivi più accreditati attualmente, il *Peceptual Assimilation Model* (PAM) di Best (1995) possa essere applicato a parlanti con una lunga formazione scolastica di L2: il fine è di capire se questa tipologia di apprendenti rientra nella categoria dei *naïve listeners* di L2 o meno.

La capacità di articolare i fonemi della L2 da parte degli studenti è stata testata attraverso l'analisi acustica delle loro produzioni mentre la capacità di percezione dei fonemi non nativi è stata testata attraverso la somministrazione di due test percettivi, l'*identification test* e l'*oddity discrimination test* (cfr. Flege & MacKay, 2004; Tsukada *et al.*, 2005). Secondo i risultati le SU sono state in grado di creare un'interlingua articolata e complessa, in cui sono emerse categorie fonetiche per ciascuno dei fonemi dell'AE. Inoltre, le categorie dell'interlingua sembrano differire dai fonemi nativi quindi è possibile affermare che le SU siano state in grado di sfruttare aree dello spazio acustico prima inutilizzate. Infine, due categorie fonetiche, /ʌ/ ed /ʊ/, sembrano essere prodotte in maniera nativa, per lo meno in termini di valori formantici medi, in quanto sono state prodotte con valori statisticamente equivalenti a quelli delle parlanti native dell'AE.

La capacità percettiva è stata analizzata attraverso l'*identification test* e l'*oddity discrimination test*, che hanno evidenziato come non tutti i fonemi di L2 generano la stessa difficoltà nell'essere discriminati: infatti, sebbene solo per alcuni contrasti, le SU hanno ottenuti risultati pari a quelli delle parlanti native di L2, dimostrando di avere una capacità di discriminazione nativa.

Nel complesso, i nostri risultati dimostrano che gli studenti discriminano i vari contrasti di L2 secondo le predizioni dal PAM, che sembrerebbe applicabile quindi anche a questa

[1] Il presente lavoro è stato concepito insieme dai due autori, tuttavia ai fini accademici sono da attribuire a Bianca Sisinni i paragrafi 2., 4., 5., 6., 7., 8. e a Mirko Grimaldi i paragrafi 3. e 9. Il paragrafo 1. è da attribuire a entrambi gli autori.

tipologia di apprendenti, i quali si comportano percettivamente come *naïve listeners* nonostante abbiano un lungo background di L2 esclusivamente scolastico.

Infine la comparazione dei dati in produzione con quelli in percezione, in particolare con i dati ottenuti nell'*identification test*, ci ha portato ad elaborare un' ipotesi che vede la comparazione acustico-percettiva sistematica fra gli spazi acustici della L1 con quelli della L2 come un momento importante dell'individuazione di salienze percettive fra contrasti fonologici. Da qui l'apprendente partirebbe per costruire l'abilità a discriminare i suoni della L2 e ad estrarre dal segnale acustico tratti articolatori invarianti per generare una rappresentazione mentale astratta di un set di categorie fonetiche prima e fonologiche poi. La fase finale di questo processo porterà gli apprendenti a proiettare adeguatamente tali rappresentazioni motorie astratte sull'apparato fonatorio nella produzione dei suoni della L2. Queste ipotesi sono state brevemente discusse rispetto ai modelli percettivi *top-down* e ad alcuni risultati della recente letteratura neurocognitiva.

2. INTRODUZIONE

In letteratura si è discusso ampiamente del rapporto che intercorre fra la percezione e la produzione dei fonemi non nativi da parte di parlanti nativi di L1. L'idea prevalente, sia pure non da tutti accettata,[2] ritiene che per spiegare appieno l'acquisizione fonologica di una L2 bisogna prima spiegare il modo in cui i parlanti della L2 riescono a sviluppare una percezione appropriata e quindi una rappresentazione cognitiva dei segmenti della L2: la produzione corretta sarebbe una diretta conseguenza della corretta rappresentazione astratta. Se ne deduce che la percezione precederebbe la produzione dei fonemi di L2 (Bion *et al.*, 2006), e l'accuratezza con cui questi vengono articolati sembra essere vincolata all'accuratezza con cui vengono percepiti (Flege *et al.*, 1999). A sua volta, la capacità di percezione sembra dipendere fortemente dal grado di similarità/dissimilairtà degli inventari fonologici del sistema della L1 rispetto a quello della L2, che potrebbe essere uno dei fattori principali coinvolti nell'acquisizione delle categorie fonologiche non native. Attualmente due modelli teorici si propongono di studiare la percezione e la produzione dei fonemi di L2 da parte di parlanti nativi di L1, il *Perceptual Assimilation Model* (PAM; Best, 1995) e lo *Speech Learning Model* (SLM; Flege, 1995).

Partendo da questi presupposti, gli obiettivi del presente lavoro possono essere così sintetizzati:

(a) valutare le abilità di parlanti adulti dell'IS di percepire ed articolare i fonemi dell'AE (L2).
(b) cercare di comprendere quale sia la relazione effettiva che intercorre fra il processo di percezione e quello di produzione dei fonemi non nativi.
(c) verificare se il modello PAM, che spiega come *naïve listeners* discriminano contrasti fonetici non nativi, si possa applicare agli studenti universitari, e di conseguenza capire se anche questa tipologia di apprendenti si comporta come *naïve listeners* di L2 nonostante una lunga carriera scolastica alle spalle.

[2] Per una discussione più approfondita della questione ed un'analisi da un punto di vista neurofisiologico cfr. Grimaldi *et al.* (in stampa).

2.1. I modelli teorici sull'acquisizione dei fonemi di L2

Gli studi sull'acquisizione della L2 sono stati spesso condotti su tipologie di soggetti differenti. Si va da *highly experienced learners* – ovvero soggetti esposti sin dalla prima infanzia e nel paese straniero alla L2 (con un elevata *Age of Arrival*, AOA) – e residenti nel paese di L2 per lungo tempo (*length of residence*, LOR), i quali usano la L1 raramente rispetto all'uso che fanno di L2 (Flege, 1995; Flege, 1997; Flege *et al.*, 1999), ad apprendenti di L2 in contesti scolastici aventi un livello di conoscenza avanzato (Flege & MacKay, 2004; Lengeris & Hazan, 2007; Mora & Fullana, 2007). Queste due differenti tipologie di parlanti riflettono a loro volta due differenti condizioni di acquisizione/apprendimento della L2, ovvero la *Second Language Acquisition* (SLA) e la *Classroom Foreign Language Acquisition* (FLA). L'acquisizione della L2 ha luogo in ambito SLA quando si è completamente calati nel contesto della L2 e questa lingua diventa di uso predominante. Contemporaneamente, l'uso della L1 viene confinato a situazioni limitate come, ad esempio, il contesto familiare. L'acquisizione/apprendimento della L2 in ambito FLA avviene invece nel contesto scolastico, in cui la lingua straniera è prevalentemente usata da insegnati non madrelingua, e dove l'uso della L1 resta, nel complesso, predominante. In sostanza, l'esposizione alla L2 in ambito SLA avviene attraverso le interazioni quotidiane con i parlanti nativi di L2 con intenti comunicativi e pragmatici; invece in ambito FLA l'esposizione alla L2 avviene attraverso l'istruzione formale con un insegnamento incentrato, fondamentalmente, sulla grammatica e sul lessico. Tuttavia, l'ambito FLA può essere considerato come un importante punto di riferimento e come valido elemento di confronto con gli studi condotti in ambito SLA (Best & Tyler, 2007).

Allo stato attuale, i modelli teorici elaborati per cercare di descrivere i processi di acquisizione dei fonemi non nativi in relazione alle modalità con cui si verifica l'esposizione alla L2 sono sostanzialmente due: lo SLM e il PAM. Il primo modello si propone di studiare la produzione dei fonemi non nativi da parte degli *high experienced learners*, mentre il secondo modello si prefigge di studiare la percezione di foni non nativi da parte dei *naïve listeners*. Questi ultimi sono parlanti monolingui che non usano attivamente la L2 e che sono totalmente estranei ad essa. In questa tipologia possono rientrare anche coloro che hanno avuto un'esposizione passiva alla L2 o limitata a contesti istituzionali con insegnanti con un forte accento nativo. Conseguentemente, sembra possibile affermare che il PAM possa essere applicabile anche in contesti formali, per lo meno in alcuni casi. Una recente estensione del modello PAM è il PAM-L2 (Best & Tyler, 2007), che si prefigge di descrivere come i sistemi fonetico-fonologici di L1-L2 interagiscono nel primo stadio di acquisizione di L2. Questo modello formula delle predizioni sulla formazione di categorie fonologiche per i foni di L2: in generale, gli apprendenti tardivi di L2 discrimineranno meglio i contrasti fonetici che deviano maggiormente dalle categorie fonologiche native in termini di parametri gestuali e, con l'aumentare dell'esperienza con la L2, costituiranno per essi nuove categorie fonetiche e fonologiche. Gli apprendenti tardivi di L2 avranno, per questi stessi contrasti, un livello di discriminazione maggiore rispetto a quello dei parlanti monolingui nativi di L1 (Avesani *et al.*, 2008).

Lo SLM ed il PAM condividono alcuni presupposti teorici e, al contempo, differiscono in altri. L'assunto principale dello SLM è relativo alla capacità di acquisizione della L2 che resterebbe intatta nell'arco dell'esistenza permettendo anche a parlanti adulti o al di fuori del Periodo Critico (Lenneberg, 1967) di formare nuove categorie fonetiche per i fonemi di

L2. [3] Le categorie fonetiche, secondo Flege (1995), sono delle rappresentazioni mentali dei suoni di L2 che si collocano nella memoria a lungo termine e che coesistono con i fonemi nativi in uno spazio fonologico comune. La coesistenza dei sistemi fonologici genererebbe l'interazione fra gli stessi e ciò avverrebbe attraverso due meccanismi: la *Category Assimilation* e la *Category Dissimilation.* Il primo meccanismo avrebbe luogo quando un parlante nativo di L1, percependo un fonema non nativo come molto simile ad un vicino fonema nativo, non riesce a creare una nuova categoria fonetica per il fonema di L2 e crea una categoria intermedia (*merged category*) fra quest'ultimo ed il più vicino fonema di L1. Questa categoria fonetica prodotta con un *undershoot* formantico, ha le caratteristiche acustiche di entrambi i fonemi e viene utilizzata dal parlante sia nella sua L1 che nella sua L2. Il secondo meccanismo, al contrario, si realizzerebbe quando il parlante nativo di L1, percependo un fonema di L2 come differente e nuovo rispetto ai fonemi nativi, riesce a formare una nuova categoria fonetica per esso che, tuttavia, differirà anche dalla categoria fonetica prodotta dai parlanti nativi di L2. Nello specifico, questa nuova categoria fonetica di L2 viene collocata, attraverso un *overshoot* formantico, in una porzione dello spazio acustico differente da quella in cui è collocato il corrispettivo fonema non nativo; inoltre, il fonema nativo più vicino viene allontanato dalla posizione originaria per preservare il contrasto fonologico fra i sistemi di L1 ed L2 (Flege *et al.*, 2003). La creazione di una nuova categoria fonetica con l'aumentare dell'età, però, diventa per i parlanti nativi di L1 un compito sempre più arduo in quanto ad un AOA elevato corrisponde una sempre maggiore difficoltà nel percepire i fonemi non nativi come differenti dai fonemi nativi.

Anche secondo il PAM, come per lo SLM, gli apprendenti adulti di L2 preservano la capacità di acquisire fonemi di L2, riuscendo ad affinare le loro abilità percettive con l'aumentare della esperienza con la L2. A differenza dello SLM, per il PAM la categoria fonetica non si identifica con rappresentazioni mentali generate dall'individuazione passiva di tratti acustico-fonetici, ma si identifica con un set di articolazioni del tratto vocale. Tale categoria fonetica differisce dalla categoria fonologica che invece distingue coppie lessicali ed i loro significati. Le categorie fonologiche sono rappresentate da quei suoni che distinguono coppie minime (/pazza/ vs. /pizza/), mentre le categorie fonetiche, pur essendo categorizzabili con uno stesso fonema, hanno una realizzazione acustico-articolatoria differente, ad esempio, fra differenti varietà regionali ([kasa]-[kaza]).

Inoltre, per lo SLM l'identificazione dei fonemi di L2 ai fonemi di L1 viene effettuata a livello fonologico e fonetico mentre, per il PAM, l'identificazione dei fonemi avverrebbe a livello fonetico (per i *naïve listeners* i fonemi di L2 sono foni della loro L1). Per il PAM-L2, invece, l'assimilazione dei foni avverrebbe a livello fonologico ma non necessariamente a livello fonetico (cfr. Best & Tyler, 2007 per una comparazione più dettagliata fra i due modelli teorici).

[3] Lenneberg (1967) ipotizza che il linguaggio umano è una capacità cognitiva che si acquisisce in maniera 'normale' durante il Periodo Critico (0-12 anni). Dopo questo termine, il linguaggio può essere acquisito con difficoltà o con processi cognitivi differenti da quelli che si attivano durante l'acquisizione naturale, in quanto il cervello perderebbe la sua plasticità neurale in concomitanza della maturità. Lenneberg ha formulato questa ipotesi relativamente all'acquisizione della L1 in bambini con patologie del linguaggio. La stessa ipotesi è stata, poi, estesa all'acquisizione della L2. Tuttavia, essa è stata spesso modificata o del tutto confutata.

Nonostante queste divergenze e le differenti tipologie di soggetti a cui entrambi i modelli si riferiscono, essi condividono un importante assunto, ovvero la necessità di misurare empiricamente la distanza fonetica fra L1 e L2 effettivamente percepita dai parlanti nativi di L1 (*perceived phonetic distance*, cfr. Guion *et al.*, 2000) attraverso la somministrazione di test percettivi cross-linguistici (*identification test*). Per lo SLM la distanza fonetica sarebbe alla base dell'acquisizione di nuove categorie fonetiche e della loro produzione, mentre per il PAM essa sarebbe fondamentale nel determinare i differenti gradi di discriminazione dei contrasti fonetici non nativi.

2.2. Una panoramica degli studi sulla produzione e percezione dei fonemi non nativi

Numerosi studi si sono occupati della produzione di L2 da parte di parlanti nativi di L1. Bohn & Flege (1991), per esempio, hanno studiato la produzione dei fonemi di L2 Inglese da parte di due gruppi di parlanti nativi del tedesco che differivano fra loro rispetto agli 'anni di esperienza' di L2: il gruppo di 'inesperti' viveva nel paese di L2 da meno di 1 anno ca. (0.6 anni), mentre il gruppo di 'esperti' da circa 7.5 anni. La capacità di produzione è stata verificata attraverso la comparazione dei dati acustici da loro prodotti con i dati prodotti da un gruppo di parlanti nativi di L2. Inoltre, ad un gruppo di parlanti nativi di L2 è stato fatto eseguire un *intellegibility test* in cui si chiedeva di valutare le produzioni dei parlanti non nativi. I risultati hanno confermato l'ipotesi iniziale dello studio: ovvero il fattore 'esperienza di L2' non sembra aver inciso sulla produzione dei fonemi non nativi simili ai fonemi nativi. Infatti, entrambi i gruppi di parlanti nativi del tedesco non sembrano differire nell'articolazione dei fonemi dell'inglese /ɛ/, /ɪ/, /iː/ simili ai corrispettivi fonemi nativi. Al contrario, il fattore 'esperienza' sembra aver avuto un'importante influenza sulla produzione del fonema /æ/ considerato come 'nuovo' rispetto all'inventario fonologico nativo, in quanto i parlanti 'esperti' di L2 sono riusciti ad articolarlo significativamente meglio dei parlanti 'inesperti'. Tuttavia, questo risultato è stato confermato dall'analisi acustica dei dati piuttosto che dall' *intellegibility test.*

Il fattore 'esperienza di L2' sembra aver giocato un ruolo simile anche in uno studio successivo (Flege *et al.*, 1997a), in cui è stata studiata anche la percezione dei fonemi di L2. Anche in questo caso il gruppo di 'esperti' ha eseguito *performance* migliori del gruppo di 'inesperti'. Un risultato interessante emerso da questo lavoro è che la capacità di percezione e produzione dei fonemi di L2 sembra dipendere dalla relazione fra gli inventari fonologici dei sistemi nativi e non nativi.

In Flege *et al.* (1997b) è stato studiato un altro fattore ritenuto estremamente importante dagli autori, ovvero l'uso di L1. Le produzioni di L2 inglese di due gruppi di parlanti nativi dell'italiano sono state valutate da parlanti nativi di L2. I gruppi di parlanti nativi dell'italiano avevano lo stesso AOA ma differivano nell'uso che facevano della loro L1. I risultati hanno dimostrato che una maggiore attestazione delle categorie di L1 influenza fortemente la produzione di L2 in quanto le produzioni del gruppo di Italiani che parlavano spesso la loro L1 erano stato giudicate come aventi un forte accento straniero rispetto a quelle del gruppo che utilizzava prevalentemente la L2 sulla L1.

Un risultato simile è stato ottenuto anche da Piske *et al.* (2002) che hanno osservato come le produzioni di parlanti con un'elevata esperienza di L2 ed un uso scarso della L1 hanno ottenuto gli stessi *rate* delle produzioni dei parlanti nativi di L2, mentre le produzioni di parlanti esperti di L2 ed un uso elevato della L1 e le produzioni dei parlanti con poca esperienza di L2 sono state giudicate con *rate* significativamente più bassi rispetto a quelli dei parlanti nativi di L2.

In Flege *et al.* (1999), inoltre, è stato studiato il rapporto che intercorre fra le capacità di percepire ed articolare la L2 ed è stato osservato che quanto meglio i fonemi di L2 venivano percepiti, tanto meglio venivano prodotti, portando alla conclusione che la percezione sembra precedere la produzione dei fonemi non nativi.

Con lo studio di Flege *et al.* (2003) si conferma l'importanza dell'esperienza di L2 e dell'uso che si fa della propria lingua materna. Quattro gruppi di parlanti nativi dell'italiano e di L2 inglese, differenziati per esperienza ed uso della L2 – *high experienced-low L1 use*; *high experienced-high L1 use*; *low experienced-low L1 use*; *low experienced-high L1 use* – hanno prodotto il fonema /eɪ/ dell'Inglese. Solamente il gruppo *high experienced-low L1 use* è riuscito a costituire una nuova categoria fonetica per il fonema /eɪ/ di L2, che è stato prodotto con un maggiore 'spostamento' formantico (*overshoot*) rispetto a quello effettuato dai parlanti nativi dell'inglese. Al contrario, i restanti gruppi non sono riusciti a costituire una nuova categoria fonetica e hanno prodotto il fonema non nativo con un *undershoot* formantico rispetto a quello della stessa categoria prodotta dai parlanti nativi. Hanno creato, quindi, una sorta di *merged category* fra il fonema non nativo ed il più vicino fonema nativo.

Gli studi passati in rassegna possono riferirsi agli *L2 learners*, ovvero a parlanti nativi di L1 con un'elevata conoscenza di L2 sviluppata in un contesto naturale in cui hanno vissuto per diversi anni. In Guion *et al.* (2000) sono state invece comparate le *performance* percettive di tre gruppi di parlanti nativi del giapponese differenti in 'esperienza di L2' inglese: si va dal gruppo che aveva una conoscenza scolastica (*naïve listeners*) al gruppo che viveva nel paese di L2 da 3 anni (LOR). Gli autori hanno osservato che i tre gruppi non differivano fra loro nel discriminare i contrasti fonologici dell'inglese e che il loro livello di discriminazione era significativamente più basso rispetto di quello dei parlanti nativi di L2. Inoltre, le capacità di discriminazione dei tre gruppi sono risultate in linea con le predizioni del PAM (cfr. 9.2), per cui anche coloro con un LOR pari a 3 anni si comportano percettivamente come *naïve listeners* dell'inglese. Gli autori giungono alla conclusione che il PAM può essere applicato anche nelle prime fasi di acquisizione della L2 in contesto naturale.

3. ITALIANO SALENTINO E INGLESE AMERICANO

Prima di entrare nel vivo di questo studio, riteniamo sia utile effettuare un confronto preliminare fra i due sistemi fonologici che saranno oggetto di analisi: il sistema dell'IS e quello dell'AE. L'IS è la varietà di italiano parlata nel Salento, e risente del sistema fonologico dialettale proprio di un'area della Puglia che si estende dall'antica Via Appia sino al Capo di Santa Maria di Leuca, e che amministrativamente è circoscritta dalle province di Lecce, Brindisi e dalla parte meridionale della provincia di Taranto.

L'inventario fonologico dell'IS è costituito dai 5 fonemi /i/, /E/, /a/, /O/, /u/ (cfr. Grimaldi, 2003, 2009),[4] a differenza dell'italiano standard che annovera i 7 fonemi /i/, /e/, /ɛ/, /a/, /ɔ/, /o/, /u/ (Grassi *et al.*, 1997): l'IS, infatti è privo dell'opposizione fonematica fra le vocali medio-alte e medio-basse anteriori e posteriori (/e/-/ɛ/ e /ɔ/-/o/), quindi i gradi di apertura di questa varietà sono tre (alto, medio, basso) e non cinque.

[4] La particolare notazione fonetica di /E/ ed /O/ si riferisce al fatto che le analisi acustiche hanno evidenziano che si tratta di fonemi né chiusi né aperti, ma collocati in posizione intermedia fra /i/ ed /a/ sull'asse anteriore ed /u/ ed /a/ su quello posteriore.

L'AE, più specificamente il *General American English* parlato negli USA ed in Canada,[5] presenta un inventario fonologico costituito da 10 fonemi vocalici non dittongati /iː/, /ɪ/, /ɛ/, /æ/, /ʌ/, /ɑː/, /ɝː/, /ɔː/, /ʊ/ e /uː/ (Ladefoged, 2001) con cinque gradi di apertura (alto, medio-alto, medio, medio-basso, basso). Alcuni di questi fonemi sono definiti dalla presenza del tratto soprasegmentale di durata e, infatti, molto spesso viene presa in considerazione la distinzione fra fonemi lunghi e fonemi brevi. Tuttavia bisogna tenere presente che tale distinzione è alquanto relativa poiché la lunghezza dei fonemi vocalici dell'AE varia a seconda di numerosi fattori, fra cui i contesti consonantici e la qualità della sillaba in cui sono inseriti.[6]

Un ulteriore elemento che caratterizza questa lingua è la presenza del fonema /ɝː/: esso è uno dei fonemi più rari al mondo, ma nell'AE è estremamente comune. Ladefoged (2001) lo definisce come 'r-coloured' per il connubio del suono vocalico con il fonema consonantico /r/.

È facile osservare, dunque, le differenze che intercorrono fra i due sistemi fonologici, come i diversi gradi di apertura, la presenza/assenza del tratto soprasegmentale di durata e le presenza del fonema rotacizzato. Dati questi presupposti risulta di notevole interesse osservare come i parlanti nativi dell'IS processino i fonemi dell'AE sia in fase di percezione che di produzione.

4. METODO

4.1. Partecipanti

In questo studio sono state prese in esame 18 studentesse universitarie (SU) – età media 20,4 anni frequentanti il primo anno di corso di inglese tenuto da una lettrice madrelingua AE presso la Facoltà di Lingue dell'Università del Salento. In un questionario di autovalutazione dai noi fornito, le SU hanno riportato di aver cominciato a studiare inglese come L2 all'età media di 10.3 anni e di non essere mai state all'estero per un periodo superiore ad un mese. Delle 18 SU totali, 14 di loro hanno affermato di aver avuto un insegnante madrelingua inglese durante il corso delle scuole medie superiori, mentre tutte asseriscono di aver avuto un insegnamento principalmente incentrato sull'apprendimento della grammatica. Ognuna delle SU è stata sottoposta ai test di produzione e percezione nella stanza insonorizzata del CRIL per la durata di 1h circa.

4.2. Materiale audio

Gli stimoli audio utilizzati per i test sono stati prodotti da 3 parlanti native dell'AE (AES). Ogni fonema dell'AE (/iː/, /ɪ/, /ɛ/, /æ/, /ʌ/, /ɑː/-/ɑ˞ː/, /ɝː/, /ɔ˞ː/, /ʊ/, /uː/) è stato fatto realizzare in parole monosillabiche con eguale contesto consonantico /p_t/ (a parte il fonema /uː/, che, per mancanza di forme lessicali, è stato fatto produrre nel contesto /b_t/, con la consonante iniziale che differisce solo per tratto di sonorità). Le parole sono poi state inserite nella frase cornice "I say /p_t/ now", seguite dalla frase-stacco "Could you

[5] In questo lavoro si trascurano le differenti varietà dell'inglese Americano e si prende come riferimento la varietà standard che "is [also] used as a model by millions of students learning English as a second language" (Collins & Mees, 2003: 6).

[6] Strange *et al.* (1998) affermano che le vocali cosiddette lunghe, quando inserite negli stessi contesti, hanno una durata effettivamente maggiore del 30-45% rispetto alla durata delle vocali brevi.

repeat, please" per 6 volte (cfr. 6.1). Ogni AES ha prodotto gli stimoli necessari per i test all'interno della stanza insonorizzata del CRIL. Le produzioni sono state registrate con CSL 4500 ad una frequenza di campionamento di 22.05 Khz, e successivamente sono state segmentate e normalizzate in intensità con Praat 4.6.29.

5. TEST DI PRODUZIONE

5.1. Elicitazione dei dati

Le produzioni delle SU per la seconda lingua, sono state elicitate facendo ricorso alla cosiddetta *delayed repetition technique* (Flege, 1995). Tale tecnica consiste nel far leggere ai soggetti una frase-cornice – "I say /p_t/ now" – sullo schermo del PC, mentre, nel contempo, viene fatta sentire la stessa frase prodotta alternativamente da ognuna delle 3 AES. In questo modo si cerca di ovviare alla possibile influenza dello spelling dell'AE sulle produzioni delle SU combinando, con lo stimolo visivo, lo stimolo uditivo 'guida'. Come anticipato, alla frase-cornice prodotta da una delle 3 AES è stata fatta seguire un frase-stacco – "Could you repeat, please?" – e due *bip* distrattori (1.3 sec ciascuno) alternati con due pause di silenzio (3 sec. ciascuna), per un totale di 8.6 sec. Inoltre, alle SU è stato chiesto di leggere un'ulteriore frase-stacco "Of course I could", prima di leggere e ripetere la frase-cornice contenente la parola bersaglio, in modo tale da prolungare ulteriormente il *delay* fra l'ascolto del fonema *target* e la sua ripetizione e per ridurre ulteriormente il rischio di imitazione.[7]

In totale, le SU hanno prodotto 66 fonemi dell'AE, poiché per ogni fonema sono state scelte due frasi-cornice (seguite dalle frasi-stacco) prodotte da ciascuna delle 3 AES. In sostanza, durante l'esperimento ogni SU ha simulato il mini-dialogo in (1) con le parlanti native dell'AE:

(1) AES: I say ___ now. Could you repeat, please?
(PAUSA – BIP – PAUSA – BIP = durata totale 8,6 sec)
SU: Of course I could. I say ___ now

Le produzioni di ogni SU sono state elicitate prima dell'esecuzione dei test percettivi per impedire che l'esposizione prolungata agli stimoli dell'AE potesse influire in qualche modo sulla loro pronuncia. L'esigenza di comparare il livello di produzione dei fonemi della L1 rispetto a quello della L2, ci ha costretto a bilanciare i due *corpora* di rilevamento. Per quanto possibile, abbiamo inserito i fonemi *target* nativi e non nativi negli stessi contesti consonantici. Tuttavia, ciò non sempre è stato possibile, per cui nel corpus della L1 abbiamo dovuto ricorrere a delle pseudo parole. I fonemi nativi *target* sono stati inseriti nella sillaba tonica di parole bisillabiche avente gli stessi contesti consonantici dei monosillabi dell'AE (/ˈp_tta/ vs /p_t/): ad esempio, per la produzione della L2, le SU hanno prodotto i monosillabi /peat/ (/iː/), /pit/ (/ɪ/), /pet/ (/ɛ/), ecc, mentre, per la produzione dei fonemi nativi hanno realizzato pseudo- parole e parole come /ˈpitta/ (/i/), /ˈpetta/ (/E/), ecc.

Il fonema /u/ è stato inserito anche nella parola /ˈb_tta/ per renderlo paragonabile con l'AE /boot/; mentre i fonemi /a/, /E/ e /O/ sono stati inseriti anche in /ˈp_rta/ per simulare quanto più possibile la realizzazione dei fonemi rotacizzati /ɑˠː/, /ɜˠː/, /ɔˠː/. Il nucleo della sillaba atona è rappresentato sempre dal fonema /a/, che in italiano sembra essere articolato

[7] "[…] the intervening speech material probably prevented direct imitations from sensory memory" (Piske *et al*., 2001: 206).

con il tratto vocale in una posizione abbastanza neutra. Le parole bisillabiche sono state a loro volta inserite in frasi-cornice aventi la stessa struttura e lo stesso significato delle frasi-cornice di L2 ("Dico ___ adesso"). In totale sono stati creati 54 stimoli per il corpus dell'IS: 6 stimoli per ogni fonema, come per il corpus dell'AE.

5.2. Produzione di L1 ed L2: analisi acustica e statistica

Nelle Tabelle 1 e 2 sono illustrati, rispettivamente, i valori formantici medi della L1 delle SU e quelli della L2 delle SU e delle AES:

	/i/	/E/	/a/	/O/	/u/	/Er/	/ar/	/Or/	/bu/
F1	316 (18)	465 (37)	655 (33)	503 (26)	344 (25)	551 (37)	661 (32)	506 (22)	342 (18)
F2	2017 (45)	1817 (55)	1341 (48)	1128 (43)	1018 (62)	1670 (83)	1365 (45)	1161 (44)	1023 (56)

Tabella 1: Valori formantici medi dei fonemi della L1 delle SU

Voc.	L2 SU		AES	
	F1	F2	F1	F1
/iː/	345 (34)	1997 (72)	319 (16)	2141 (65)
/ɪ/	357 (32)	1972 (69)	413 (14)	1807 (44)
/ɛ/	553 (36)	1663 (93)	529 (23)	1695 (84)
/æ/	601 (43)	1528 (107)	680 (24)	1483 (41)
/ɑː/	555 (47)	1292 (93)	616 (25)	1222 (26)
/ʌ/	550 (72)	1374 (92)	574 (21)	1330 (38)
/ɑˠː/	572 (47)	1295 (93)	555 (14)	1254 (27)
/ɔˠː/	489 (37)	1188 (87)	423 (12)	1032 (40)
/ɜˠː/	516 (40)	1421 (80)	447 (42)	1367 (54)
/ʊ/	402 (55)	1232 (126)	419 (13)	1261 (37)
/uː/	365 (28)	1186 (143)	337 (14)	1139 (89)

Tabella 2: Valori formantici medi dei fonemi della L2 delle SU e delle AES

In 5.1 abbiamo visto che l'esigenza di un confronto coerente fra il sistema fonologico della L1 e il sistema in fase di apprendimento della L2 ci ha portato a strutturare il corpus di rilevamento della lingua nativa in modo da poter essere parametrizzato con le produzioni della L2. Ciò ha portato ad inserire il fonema /u/, oltre che nella pseudo-parola /ˈp_tta/, anche nella parola /ˈb_tta/ per un confronto con /boot/ dell'AE, e i fonemi /a/, /E/, /O/ nella pseudo-parola /ˈp_rta/ - oltre a /ˈp_tta/ - per una comparazione con i fonemi rotacizzati /ɑˠː/, /ɜˠː/, /ɔˠː/. Per verificare se la produzione dei fonemi dell'IS nella parole e pseudo-parole /ˈb_tta/ vs /ˈp_tta/ (per /u/) e /ˈp_rta/ vs /ˈp_tta/ (per /a/, /E/, /O/) possa essere stata condizionata dal particolare contesto in cui sono stati inseriti, abbiamo eseguito una serie di T-test a campioni indipendenti. Quindi sono stati comparati fra loro i valori di F1 ed F2 di /u/ in /ˈb_tta/ e /ˈp_tta/ che sono risultati statisticamente equivalenti (F1 e F2 $p > 0{,}01$): data questa equivalenza, nelle fasi successive di questo lavoro sono stati presi in considerazione i soli valori di /u/ preceduta da /p/.

La stessa serie di test statistici è stata eseguita per /a/ rispetto al contesto /ar/, per /E/ rispetto ad /Er/, e per /O/ rispetto ad /Or/. In questo caso è emerso che i tre fonemi

presentano valori formantici che differiscono significativamente fra loro, in funzione della presenza o meno della vibrante /r/ (/a/-/ar/ F1: $p > 0,01$, F2: $p < 0,01$;/E/-/Er/ F1: $p < 0,01$, F2: $p < 0,01$; /O/-/Or/ F1: $p > 0,01$, F2: $p < 0,01$). In particolare, /E/ nel contesto /Er/ subisce un abbassamento e una leggera centralizzazione (dovuta alla diminuzione dei valori di F2), mentre /a/ nel contesto /ar/ ed /o/ nel /Or/ subiscono solo una leggera centralizzazione. Di conseguenza, è stato tenuto in considerazione il fatto che i fonemi /a/, /E/, /O/ dell'IS subiscono una variazione acustico-articolatoria quando sono seguiti da vibrante (cfr. Figura 1).

Sul versante dell'AE, poiché per il fonema /ɑː/ è stata presa in considerazione anche la forma rotacizzata /ɑ˞ː/ abbiamo ritenuto opportuno effettuare la stessa serie di T-test a campioni indipendenti. Anche in questo caso, i valori formantici fra il fonema /ɑː/ e la sua controparte rotacizzata differiscono significativamente fra loro (F1: $p > 0,01$; F2: $p < 0,01$): il fonema /ɑ˞ː/ è risultato più anteriore di /ɑː/.

Un'ulteriore analisi è stata svolta al fine di valutare in termini di significatività statistica quanto e con che modalità il sistema fonologico della L1 si differenzia da quello della L2. Per confrontare i fonemi della L1 con i fonemi dell'AE sono stati presi come riferimento i risultati ottenuti nell'*identification test* (cfr. 7. e 7.1), che ci fornisce informazioni precise su quali suoni della L2 vengono identificati con i suoni della L1 e con che percentuale. Sulla base di questi risultati (vedi Tabella 3) sono stati comparati statisticamente i seguenti gruppi di fonemi dell'AE con i fonemi dell'IS:

- /æ/, /ɑː/ ed /ʌ/ dell'AE con /a/ dell'IS: i primi due sono risultati essere significativamente differenti da /a/ sia per F1 ($p < 0,01$) che per F2 ($p < 0,01$), mentre /ʌ/ condivide con /a/ solo il tratto di anteriorità (F2: $p > 0,01$) ma non quello di altezza (F1: $p < 0,01$).
- /ɑ˞ː/ dell'AE con /ar/ dell'IS: il primo si differenzia dal secondo sia per F1 che per F2: $p < 0,01$.
- /æ/, /ɛ/ ed /ɜ˞ː/ dell'AE con /E/ ed /Er/ dell'IS: tutti e tre differenti sia per F1 che per F2 ($p < 0,01$).
- /ɔ˞ː/dell'AE con /O/: differente sia per F1 che per F2 ($p < 0,01$).
- /iː/, /ɪ/ dell'AE con /i/ dell'IS: il primo condivide con il fonema dell'IS il tratto di altezza (F1: $p > 0,01$; F2: $p < 0,01$), mentre il secondo è totalmente differente (F1 e F2: $p < 0,01$).
- /uː/ e /ʊ/ dell'AE con /u/ dell'IS: condivide con il fonema dell'IS il tratto di altezza (/uː/-/u/: F1: $p > 0,01$; F2: $p < 0,01$); mentre il secondo si differenzia del tutto F1 e F2: $p < 0,01$).

Una agevole comparazione fra il sistema fonologico dell'IS e quello dell'AE può essere ottenuta sovrapponendo su una piano cartesiano, con la F1 in ascissa e la F2 in ordinata, le aree di esistenza in Hz-like dei fonemi prodotti dalle SU e dalle AENS, come si può vedere in Figura 1. Con lo stesso metodo possiamo invece ottenere le aree di esistenza dei fonemi della L2 prodotti dalle SU, come esemplificato in Figura 2.

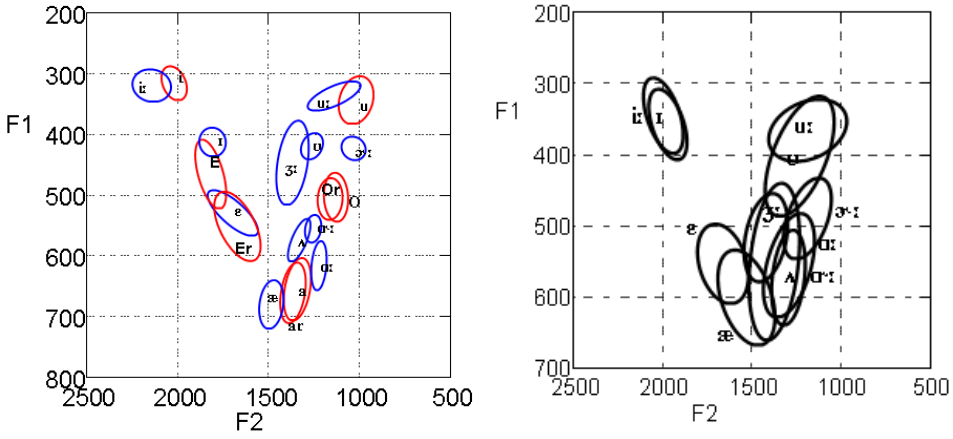


Figure 1-2: Aree di esistenza dei fonemi dell'AE (blu) e dell'IS (rosso) sulla sinistra ed aree di esistenza della L2 delle SU sulla destra

6. TEST PERCETTIVI: *IDENTIFICATION TEST* E *ODDITY DISCRIMINATION TEST*

Lo scopo dell'*identification test* è quello di misurare la 'distanza fonetica' fra i suoni di due lingue, ovvero di individuare i suoni della L2 che vengono considerati più o meno simili ai suoni della L1 e che, di conseguenza, sono più o meno difficili da discriminare. Abbiamo visto come il *Perceptual Assimilation Model* (PAM, Best, 1995) e lo *Speech Learning Model* (SLM, Flege, 1995) sostengono, seppur con prospettive differenti, che la distanza fonetica, o meglio il grado di similarità/dissimilarità fra i suoni di L1 e di L2 sia alla base della percezione di questi ultimi (Best & Tyler, 2007; Flege, 1995). L'*identification test* misura questa distanza e permette di individuare coppie di suoni di L2 più o meno facili da discriminare, grazie ad ulteriori informazioni ottenute attraverso il *rating* assegnato ai singoli stimoli percepiti da parte dei soggetti. Apposite procedure di misurazione statistica del *rating* fornito dai soggetti (vedi oltre) portano all'identificazione di coppie di contrasti più o meno difficili da discriminare, che in un secondo momento saranno testati nell'*oddity discrimination test*.

Le SU sono state testate con delle cuffie che permettono la regolazione del volume a piacimento. Gli stimoli per i due test sono stati ricavati dalle produzioni delle *frasi cornice* delle AES con l'aggiunta di stimoli nella cornice consonantica /s_t/ per i fonemi /iː/, /ʊ/ e /uː/, per un totale di 36 stimoli uditivi (12 fonemi dell'AE x 3 AES). Le parole sono state normalizzate in intensità e segmentate con Praat 4.6.29. I 36 stimoli uditivi sono stati presentati tramite computer per 3 volte con un ordine casuale: le SU li dovevano identificare ed associare alle 5 vocali dell'italiano, cliccando su ciascuna di esse. Il compito successivo richiedeva che i soggetti giudicassero quanto il suono della L2 appena sentito potesse essere identificato con il suono di L1 da loro prescelto (*goodness of fit*, GOF), cliccando su un valore compreso tra 1 (per niente identificabile) e 5 (totalmente identificabile). Moltiplicando il GOF per le percentuali di identificazione di ciascun fonema, si

ottiene il *fit index*, un indice fortemente indicativo di come il fonema non nativo sia stato effettivamente identificato con un determinato fonema nativo (Guion *et al.*, 2000).

Prima di eseguire il test, alle SU sono state fornite istruzioni orali ed è stato fatto eseguire un pre-test (con 10 stimoli) con la supervisione dello sperimentatore, per accertarsi che le istruzioni fossero state comprese e che il compito venisse eseguito correttamente.

6.1 Identification test

I risultati ottenuti da 1944 giudizi (dati dalle 18 SU X 36 stimoli X 3 ripetizioni) hanno condotto all'individuazione di 7 contrasti più o meno difficili da discriminare, ovvero /iː/-/ɪ/, / ɛ/-/ɜˑː/, /æ/-/ɑː/, /æ/-/ʌ/, /ɑː/-/ʌ/, /ɑˑː/-/ɔˑː/ e /uː/-/ʊ/, e di un contrasto, al contrario, molto semplice da discriminare, /iː/-/uː/. Quest'ultimo è stato quindi utilizzato come contrasto di controllo nel secondo test percettivo, l'*oddity discrimination test*. Nella Tabella 3 sono illustrate le percentuali di identificazione, il GOF (fra parentesi) di ciascun fonema dell'AE rispetto ai fonemi dell'IS e i *fit index* (FI), (Guion *et al.*, 2000).

L1	/i/		/E/		/a/		/O/		/u/	
L2	Perc./GOF	FI	Perc./GOF	FI	Perc./GOF	FI	Perc./GOF	FI	Perc./GOF	FI
/iː/	100% (4,4)	4,4								
/ɪ/	99% (4,1)	4								
/ɛ/			99% (4,1)	4						
/æ/			58% (3,5)	2	42% (3,2)	1,3				
/ʌ/			3% (2)		67% (3,1)	2	20% (3)	0,6	8% (3,3)	
/ɑː/					89% (3,6)	3,2	10% (3,1)	0,3		
/ɜˑː/			74% (2,1)	1,5	9% (2,2)	0,1	11% (2,4)	0,2	6% (2,5)	0
/ɑˑː/			3% (1)	0	92% (3,2)	2,9	5% (3)	0,1		
/ɔˑː/							100% (3,8)	3,8		
/ʊ/			2% (1)				15% (2,4)	0,3	83% (3,3)	2,7
/uː/									100% (3,9)	3,9

Tabella 3: Percentuali di identificazione e GOF (fra parentesi) dei fonemi dell'AE (L2) con i fonemi dell'IS (L1), e *fit index* (FI), indice di come il fonema della L2 viene identificato con un determinato fonema della L1

Le percentuali di identificazione risultano estremamente utili per poter stabilire in che misura i fonemi dell'AE sono stati identificati con i fonemi dell'IS, ovvero per capire se i fonemi di L2 sono stati identificati con i fonemi nativi in maniera consistente oppure no (Guion *et al.*, 2000). In questo caso, un fonema della L2 è identificato in maniera consistente con un fonema della L1 quando presenta percentuali di associazione pari o superiori al 70%, e solo la prima identificazione modale viene presa in considerazione; al contrario un fonema della L2 non è identificato in maniera consistente con un fonema della L1 quando presenta percentuali di associazione inferiori al 70%, e le prime due identificazioni modali sono state prese in considerazione.

Sulla base di questa procedura, le SU hanno identificato in maniera consistente i seguenti fonemi dell'AE con dell'IS:

- /iː/ e /ɪ/ sono stati entrambi identificati con il fonema nativo /i/: /iː/ percentuale di identificazione = 100%, *fit index* = 4,4; /ɪ/: percentuale di identificazione = 99%, *fit index* = 4).
- /ɛ/ è stato ampiamente identificato con il fonema /E/: percentuale di identificazione = 99%, *fit index* = 4.
- /ɑː/ e /ɑ˞ː/ con il fonema /a/: /ɑː/ percentuale di identificazione = 89%, *fit index* = 3,2; ed /ɑ˞ː/ percentuale di identificazione = 92%, *fit index* = 2,9.
- /ɜ˞ː/ con il fonema /E/: percentuale di identificazione = 74%, *fit index* = 1,5.
- /ɔ˞ː/ con /o/: percentuale di identificazione = 100%, *fit index* = 3,8.
- /uː/ e /ʊ/ con il fonema /u/: /uː/ percentuale di identificazione = 100%, *fit index* = 3,5; /ʊ/ percentuale di identificazione = 83%, *fit index* = 2,7.

Al contrario, alcuni fonemi non sono stati identificati in maniera consistente con nessun fonema dell'IS. Come abbiamo appena detto, considerando la soglia del 70%, quando la percentuale di identificazione è scesa sotto tale soglia, sono stati presi in considerazione le prime due identificazioni modali:

- /æ/ è stato identificato con /a/ ed /E/: /a/ percentuale di identificazione = 42%, *fit index* = 3,2; /E/ percentuale di identificazione = 58% *fit index* 3,5.
- /ʌ/ è stato identificato con /a/ e /o/: /a/ percentuale di identificazione = 67%, *fit index* = 3,1; /o/ percentuale di identificazione = 20%, *fit index* = 0,6.

Benché le percentuali di identificazione siano un ottimo indice dei processi di identificazione dei fonemi della L2 con quelli della L1, non ci danno informazioni rilevanti su quanto i fonemi non nativi siano stati giudicati dalle SU come buoni o poveri esempi dei fonemi dell'IS. Di conseguenza, la deviazione standard è stata utilizzata come criterio per stabilire ciò (Guion *et al.*, 2000). Quindi, abbiamo prima individuato il fonema di L2 associato con la percentuale di identificazione più alta ad un fonema di L1 e con un GOF più elevato. Tale fonema, infatti, può rappresentare una buona categoria fonetico-fonologica che i due sistemi linguistici condividono e con cui gli altri fonemi non nativi si possono confrontare. Il fonema dell'AE che rientra in questo quadro è /iː/, che è stato identificato con /i/ della L1 nella totalità dei casi (cfr. Tabella 3): 100% di identificazioni, e con un GOF, coincidente con il *fit index*, più elevato, pari a 4,4. Se si prende come riferimento la deviazione standard del GOF, ossia 0,8 (che nelle tabelle non compare), e la si sottrae dalla media del GOF, si ottiene il valore 3,6. La forchetta di valori 4,4-3,6 così ottenuta ci consente di capire quali fonemi della L2 sono stati percepiti con più alta probabilità come fonemi della L1. Quindi tutti quei fonemi dell'AE identificati con un *fit index* fra 4,4 e 3,6 sono stati considerati come buoni esemplari dei fonemi nativi con cui sono stati identificati, e cioè: /iː/, /ɪ/, /ɛ/, /ɔ˞ː/ e /uː/. Inoltre, sottraendo ancora la deviazione standard di 0,8 dalla soglia 3,6 si ottiene che i fonemi aventi un *fit index* compreso fra 3,5 e 2,8 sono stati considerati come sufficienti esempi dei fonemi dell'IS, e cioè: /ɑː/e /ɑ˞ː/. Infine quei fonemi con un *fit index* pari o inferiore a 2,7 sono stati considerati come poveri esempi dei fonemi nativi a cui sono stati associati: ovvero /æ/, /ʌ/, /ʊ/ e /ɜ˞ː/.

6.2. Oddity discrimination test

Lo scopo dell'*oddity discrimination test* è quello di misurare la capacità dei soggetti di discriminare i suoni della L2. In particolare tale prova mira a capire se il soggetto riesce a percepire delle differenze in una coppia di stimoli e quindi ad individuare lo stimolo che appartiene ad una categoria fonetico/fonologica differente: in questo modo si è in grado di verificare se il soggetto abbia acquisito o meno la categoria fonologica non nativa.

Come per l'*identification test*, anche per l'*oddity discrimination test* le SU sono state testate singolarmente in una stanza insonorizzata, tramite computer e con cuffie il cui volume è regolabile piacimento. Per ogni contrasto individuato grazie all'*identification test* sono stati eseguiti 8 *change trials* e 8 *catch trials* (trials totali: 128 X SU). I *change trials* sono costituiti da 3 items, ognuno pronunciato da una delle tre AES, e uno di questi è l'*odd item*, ovvero lo stimolo deviante da discriminare. Per ogni serie di *change trials*, l'*odd item* ricorre in posizioni alternativamente differenti (iniziale, centrale, finale) in maniera quasi bilanciata per evitare *bias* nelle risposte dovuti all'ordine della presentazione degli stimoli (Bion *et al.*, 2006). Come i *change trials*, anche i *catch trials* sono stati pronunciati dalle tre AES, in modo tale che per ogni trial risultino 3 item foneticamente differenziati ma non fonologicamente differenti. In questo modo si può testare la capacità delle SU di ignorare le differenze acustiche ma non quelle fonologiche.

Ad esempio, per testare il contrasto di controllo /iː/-/uː/, i *change trials* sono strutturati nel seguente modo:

(2) /iː/-/iː/-/uː/ – /iː/-/uː/-/iː/ – /uː/-/iː/-/iː/ – /uː/-/uː/-/iː/ – /uː/-/iː/-/uː/ – /iː/-/uː/-/uː/

I *catch trials* sono stati invece strutturati come in (3):

(3) /iː/-/iː/-/iː/ – /uː/-/uː/-/uː/

dove l'unica differenza è dovuta alla variazione fonetica prodotta dalla realizzazione dello stesso fonema da parte di tre parlanti madrelingua diversi. Le SU hanno quindi cliccato in corrispondenza della posizione dell'item che hanno percepito come differente ('1', '2', '3') o sulla voce 'nessuno' se hanno percepito tutti e tre gli items come uguali.

La capacità delle SU di discriminare i 9 contrasti – /iː/-/ɪ/, /ɛ/-/ɜˠː/, /æ/-/ɛ /, /æ/-/ɑː/, /æ/-/ʌ/, /ɑː/-/ʌ/, /ɑˠː/-/ɔˠː/, /uː/-/ʊ/ e /iː/-/uː/ – individuati attraverso il test precedente è stata valutata calcolando l'indice A' applicando la formula di Sndograss *et al.* (1985). La formula prevede il calcolo delle proporzioni di *hits* (le selezioni corrette dell'*odd item* nei *change trials*) e di *false alarms* (le selezioni errate dell'*odd item* nei *catch trials*). Un A' di 1.0 è indice di un livello di discriminazione eccellente mentre un A' di 0.5 (o inferiore) è indice della mancata discriminazione dei fonemi del contrasto non nativo proposto.

Come per l'*identification test*, anche in questo caso alle SU sono state preventivamente fornite istruzioni orali ed è stato fatto eseguire un pre-test con 5 *trials* sotto la supervisione dello sperimentatore. Inoltre, questo stesso test percettivo è stato fatto eseguire anche a 10 parlanti native dell'AE per avere un gruppo di controllo. I risultati delle SU sono stati statisticamente confrontati con i risultati ottenuti dal gruppo dell'AE.

6.3. Risultati

I risultati ottenuti (cfr. Tabella 4) dimostrano che il gruppo delle SU è stato in grado di discriminare i 9 contrasti non nativi a vari livelli: si va dal contrasto di controllo /iː/-/uː/, per il quale è stato ottenuto un A' elevato (A': 0,95) al contrasto /ɑː/-/ʌ/ discriminato con un A' al di sotto della soglia prevista (A': 0,42). Per un agevole confronto, nella Tabella 4 si possono osservare i risultati ottenuti dalle SU e dalle 10 parlanti native dell'AE:

Contrasti	**A' gruppo SU**	**A' gruppo AE**
	Risultati	*Risultati*
/iː/-/ɪ/	0.64 (0,21)	0.98 (0,02)
/ɛ/-/æ/	0.67 (0,23)	0.98 (0,01)
/ɑː/-/æ/	0.72 (0,25)	0.95 (0,04)
/ɑː/-/ʌ/	0.42 (0,23)	0.80 (0,18)
/ɑˑː/-/ɔˑː/	0.94 (0,06)	0.99 (0,01)
/æ/-/ʌ/	0.85 (0,15)	0,99 (0,1)
/ɜˑː/-/ɛ/	0.85 (0,13)	0.98 (0,03)
/uː/-/ʊ/	0.77 (0,22)	0.98 (0,02)
/iː/-/ uː/	0.95 (0,04)	0,99 (0,01)

Tabella 4: A' delle SU e delle AENS (deviazione standard fra parentesi)

Per verificare le eventuali differenze fra i due gruppi e fra i contrasti è stata effettuato un GLM univariato fra Gruppi (2) x Contrasti (9) con A' come variabile dipendente. Questa analisi ha rivelato una significatività del fattore Gruppo [$F(1,294) = 113{,}335$ $p < 0{,}01$], del fattore Contrasto [$F(8,249) = 16{,}081$ $p < 0{,}01$] e dell'interazione Gruppo X Contrasto [$F(8,294) = 4{,}631$ $p < 0{,}01$].

Al fine di verificare ulteriormente queste significatività, una ANOVA univariata è stata eseguita per ogni contrasto con Gruppo come fattore e A' come variabile dipendente. In questo caso si è cercato di capire come ogni gruppo abbia discriminato ogni singolo contrasto. Questa serie di analisi ha dimostrato che il gruppo delle SU ha discriminato in modo significativamente differente dal gruppo delle AE i seguenti contrasti: /ɑː/-/ʌ/ [$F(1,28) = 18{,}7$ $p<0{,}01$], /ɑː/-/æ/ [$F(1,28) = 7{,}9$ $p<0{,}01$], /ɛ/-/æ/ [$F(1,28) = 18{,}3$ $p<0{,}01$], /iː/-/ɪ/ [$F(1,28) = 24$ $p<0{,}01$], /uː/-/ʊ/ [$F(1,28) = 20$ $p<0{,}01$]. Al contrario i due gruppi discriminano in modo equivalente i contrasti /æ/-/ʌ/ [$F(1,28) = 7{,}2$ $p>0{,}01$], /ɜˑː/-/ɛ/ [$F(1,28) = 6{,}8$ $p>0{,}01$], /ɑˑː/-/ɔˑː/ [$F(1,28) = 5{,}1$ $p>0{,}01$], /iː/-/ uː/ [$F(1,28) = 2{,}8$ $p>0{,}01$]. Nella Figura 3 sono riportati gli A' delle SU (in blu) e delle 10 parlanti native (NS) dell'AE (in verde).

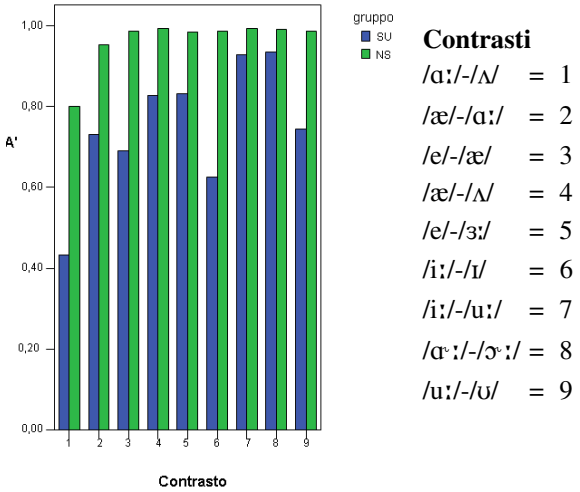


Figura 3: A' delle SU (blu) e delle 10 AENS (verde)

7. PRODUZIONE E PERCEZIONE DELLA L2 RISPETTO AI FONEMI NATIVI

Le singole analisi dei processi di produzione e percezione condotte rispettivamente in 5. e 6. lasciano però irrisolte alcune questioni, che possiamo così sintetizzare: (i) quanto le produzioni delle SU si discostano dai fonemi della L1 (e quindi si avvicinano a quelli della L2), ovvero quale livello di interlingua è stato raggiunto (se è stato raggiunto) da questi soggetti? (ii) esiste una correlazione sistematica fra i processi di produzione e quelli di percezione, oppure i due livelli sono separati, per cui quale dei due è privilegiato nel processo di acquisizione della L2?

Per rispondere a queste domande si è reso necessario dedicare una serie di test statistici alla valutazione delle capacità di articolazione dei fonemi di L2 da parte delle SU da un lato, e all'analisi incrociata delle capacità di percezione e produzione degli stessi fonemi dall'altro.

In primo luogo è stata eseguita una ANOVA univariata per comparare i valori formantici delle 3 parlanti dell'AE con i valori della L2 delle SU, e quindi verificare, in termini di valori formantici medi, quanto le SU si siano avvicinate alle produzioni delle parlanti native.

In secondo luogo si è ritenuto opportuno comparare la produzione dei fonemi della L2 delle SU con la produzione dei fonemi nativi, ricorrendo a un T-test a campioni indipendenti, sulla base dei risultati dell'*identification test*. In sostanza, ci si è proposti di capire se le dinamiche di associazione percettiva dei fonemi della L2 ai fonemi della L1 trovino una corrispondenza a livello articolatorio, o se invece le dinamiche produttive si differenzino da quelle percettive.

Infine, un'altra serie di T-test a campioni indipendenti è stata eseguita per ciascun contrasto testato nell'*oddity discrimination test*. In tal modo i valori formantici dei fonemi della L2 prodotti dalle SU sono stati comparati fra loro per verificare se esse siano state in grado di distinguere articolatoriamente categorie fonetiche non distinte percettivamente (e vice versa).

7.1. Risultati

L'ANOVA univariata con a fattore i gruppi (SU e AES) e le prime due formanti (F1 e F2) come variabili dipendenti ha messo in evidenza che le SU hanno prodotto i fonemi della L2 come di seguito indicato: /ɑː/, /ɜ˞ː/, /ɪ/, /iː/ e /ɔ˞ː/ in maniera significativamente differente dalle AES (F1 e F2: p<0,01) [/ɑː/ F1: $F(1,124) = 27,9$ p<0,01; /ɑː/ F2: $F(1,124) = 9,7$ p<0,01; /ɜ˞ː/ F1: $F(1,124) = 44,2$ p<0,01; /ɜ˞ː/ F2: $F(1,124) = 7,5$ p<0,01; /ɪ/ F1: $F(1,124) = 50,8$ p<0,01; /ɪ/ F2: $F(1,124) = 93,8$ p<0,01; /iː/ F1: $F(1,112) = 8,9$ p<0,01; /iː/ F2: $F(1,112) = 60,3$ p<0,01; /ɔ˞ː/ F1: $F(1,123) = 55,8$ p<0,01; /ɔ˞ː/ F2: $F(1,123) = 54,5$ p<0,01]; /æ/, /ɛ/ e /uː/ sono stati prodotti con i valori formantici di F2 equivalenti a quelli delle AE [/æ/ F1: $F(1,123) = 55,9$ p<0,01; /æ/ F2: $F(1,124) = 3$ p>0,01; /ɛ/ F1: $F(1,124) = 8,3$ p<0,01; /ɛ/ F2: $F(1,124) = 2,4$ p>0,01; /uː/ F1: $F(1,124) = 17,2$ p<0,01; /uː/ F2: $F(1,124) = 1,8$ p>0,01]; /ɑ˞ː/ è stato prodotto con i valori di F1 non differenti significativamente dai valori delle AES [/ɑ˞ː/ F1: $F(1,124) = 2,7$ p>0,01; /ɑ˞ː/ F2: $F(1,124) = 7,4$ p<0,01; /ʌ/ e /ʊ/ sembrano essere stati articolati in maniera nativa: i loro valori formantici medi non differiscono significativamente dai valori delle parlanti native di L2 [/ʌ/ F1: $F(1,124) = 1,9$ p>0,01; /ʌ/ F2: $F(1,124) = 3,8$ p>0,01; /ʊ/ F1: $F(1,123) = 1,7$ p>0,01; /ʊ/ F2: $F(1,123) = 0,8$ p>0,01].

Per quanto concerne il confronto fra le dinamiche in produzione e quelle in percezione, la prima serie di T-test ha messo in evidenza che le SU sono state in grado di articolare i fonemi non nativi con valori formantici differenti rispetto a quelli dei fonemi nativi a cui i primi sono stati associati nel test percettivo di identificazione (cfr. 6.1). In particolare i fonemi di L2 /iː/ e /ɪ/ differiscono dal fonema nativo /i/ [F1 di /iː/: $p < 0,01$; F2 di /iː/: $p = 0,01$; F1 di /ɪ/: $p < 0,01$; F2 di /ɪ/: $p < 0,01$]; /ɛ/ e /æ/ differiscono dal fonema dell'IS /E/ [F1 di /ɛ/: $p < 0,01$; F2 di /ɛ/: $p < 0,01$; F1 di /æ/: $p < 0,01$; F2 di /æ/: $p < 0,01$]; /ɜ˞ː/ differisce da /Er/[F1 e F2: $p < 0,01$]; i fonemi /æ/, /ʌ/ e /ɑː/ differiscono dal fonema /a/ [F1 di /æ/: $p < 0,01$; F2 di /æ/: $p < 0,01$; F1 di /ʌ/: $p < 0,01$; F2 di /ʌ/: $p < 0,01$; F1 di /ɑː/: $p < 0,01$; F2 di /ɑː/: $p < 0,01$]; il fonema /ɔ˞ː/ differisce da /Or/ [F1 e F2: $p < 0,01$]; i fonemi /uː/ e /ʊ/ differiscono dal fonema /u/ [F1 di /uː/: $p < 0,01$; F2 di /uː/: $p < 0,01$; F1 di /ʊ/: $p < 0,01$; F2 /ʊ/ di : $p < 0,01$].

La seconda serie di T-test per campioni indipendenti ha dimostrato che le SU sono state in grado di differenziare le categorie fonetiche della L2 appartenenti ad uno stesso contrasto. In particolare:

- le categorie di ciascuna delle coppie /iː/-/ɪ/, /ɛ/-/ɜ˞ː/, /ɛ/-/æ/, /æ/-/ɑː/, /æ/-/ʌ/ e /ɑ˞ː/-/ɔ˞ː/ sono state prodotte con entrambi i valori formantici differenti (F1 and F2: $p < 0,01$);
- i fonemi del contrasto /uː/-/ʊ/ sono stati prodotti con valori significativamente differenti in altezza ma uguali in posteriorità (F1: $p < 0,0,1$; F2: $p > 0,01$);
- il contrasto /ɑː/-/ʌ/ è stato articolato con valori uguali altezza ma con valori in posteriorità differenti (F1: $p > 0,0,1$; F2: $p < 0,01$).

Le AES, invece come ci si attendeva, hanno articolato i fonemi di ciascun contrasto in maniera significativamente differente in tutti i casi.

8. ABILITÀ ARTICOLATORIE E PERCETTIVE DELLE SU

Dai risultati delle analisi statistiche presentate nel paragrafo precedente si evince che le SU sono in grado di costruire delle categorie fonetiche intermedie fra le loro categorie fonologiche native e quelle della L2. Più in dettaglio, le SU sono state in grado di creare categorie fonetiche speculari a ciascun fonema dell'AE che si discostano dai loro fonemi nativi rispetto ai quali tali categorie paiono occupare varie porzioni dello spazio acustico disponibile. In sintesi, le SU sono riuscite a sfruttare, sebbene non in maniera *native-like*, porzioni di spazio acustico non utilizzate dall'inventario fonologico nativo, costituito da solo 5 fonemi rispetto ai 10 dell'AE (cfr. Figura 4).

Più in dettaglio, sulla base dei confronti statistici fra i valori formantici, notiamo che le SU sono risultate in grado di articolare le categorie fonetiche /æ/ /ɛ/ e /uː/, producendo valori di F2 che le collocano coerentemente nello spazio acustico antero-posteriore rispetto a quello della L2, ma con valori in altezza di F1 lontani da quelli della L2. La categoria fonetica /ɑˑː/, invece, è stata realizzata con valori che la collocano nella giusta centralità, ma non nella giusta altezza. Invece le categorie fonetiche /iː/, /ɪ/, /ɜˑː/, /ɑː/ e /ɔˑː/ sono state prodotte in maniera significativamente differente dalle AES sia per F1 che per F2, quindi in contrasto con il *target* articolatorio.

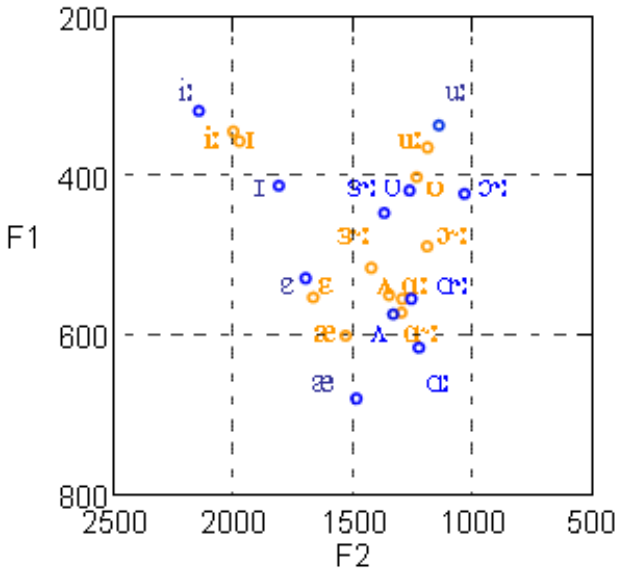


Figura 4: Valori medi della L2 delle SU (arancio) e dell'AE (blu);
i valori medi sono indicati in Tabella 2

Al contrario, le categorie fonetiche /ʌ/ e /ʊ/ sono state articolate con valori formantici statisticamente equivalenti a quelli delle corrispondenti categorie fonologiche prodotte dalle AES. Quindi è possibile supporre che, per lo meno in termini di valori formantici medi, le SU siano riuscite a produrre in maniera nativa i fonemi /ʌ/ e /ʊ/. Sebbene queste valutazioni siano state effettuate attraverso la comparazione delle categorie fonetiche nello spazio

acustico e non attraverso giudizi di intelligibilità delle stesse espressi da parlanti nativi della L2, esse sono comunque fortemente indicative delle differenze o similarità che potrebbero sfuggire all'orecchio umano (cfr. Munro, 2008).

Dalle analisi statistiche effettuate emerge un secondo elemento relativo all'interlingua delle SU: le SU, infatti, sebbene siano riuscite ad articolare solo due categorie fonetiche con valori formantici equivalenti a quelli delle SU, sono però state in grado di differenziare tutte le categorie fonetiche corrispondenti alle coppie minime dell'AE identificate con un fonema nativo o con lo stesso set di fonemi nativi (/iː/-/ɪ/, /ɜ˞ː/-/ɛ/, /ɛ/-/æ/, /ɑː/-/æ/, /æ/-/ʌ/ e /ɑ˞ː/-/ɔ˞ː/), eccetto due casi (/uː/-/ʊ/ and /ɑː/-/ʌ/). Ciò implicherebbe che le SU, nonostante abbiano articolato la quasi totalità delle categorie fonetiche della loro interlingua in maniera errata, abbiano tuttavia sviluppato una sensibilità (capacità) tale da poter permettere una differenziazione articolatoria fra fonemi non nativi riconducibili, da un punto di vista percettivo, ad una stessa o a più categorie native.

L'abilità percettiva delle SU è stata studiata attraverso l'ausilio dei due test percettivi, *identification test* e *oddity discrimination test*. Dai risultati del primo test è emerso che le SU sono riuscite a distinguere i fonemi dell'AE dai fonemi dell'IS con diversi 'gradienti' di discriminazione. Più in particolare, i fonemi dell'AE /iː/, /ɪ/, /ɛ/, /ɔ˞ː/ e /uː/ sono stati percepiti dalle SU come totalmente identificabili con i fonemi dell'IS a cui sono stati associati. I fonemi /ɑː/ ed /ɑ˞/ sono stati percepiti come assimilabili al fonema nativo prescelto, cioè ad /a/, ma non come totalmente identificabili con esso. Infine, i fonemi /æ/, /ʌ/, /ʊ/ e /ɜ˞ː/ sono stati percepiti come molto differenti dai fonemi, o dai set di fonemi, a cui sono stati associati.

Se ne deduce che non tutti i fonemi nativi della L2 sono discriminabili nello stesso modo, per cui non tutti i fonemi di L2 generano la stessa difficoltà discriminativa, visti i differenti 'gradienti' percettivi ottenuti. Inoltre, si conferma la necessità di valutare le similarità o le differenze fra due sistemi linguistici non solo da un punto di vista acustico, ma anche da un punto di vista percettivo in quanto la prossimità dello spazio acustico dei fonemi non ne implica necessariamente la similarità (Strange, 2007).

Il secondo test percettivo conferma l'ipotesi che non tutti i contrasti sono difficili da discriminare allo stesso livello. Le SU, infatti, hanno discriminato i contrasti non nativi a vari livelli, ottenendo A' pari a quelli ottenuti dalle parlanti native dell'AE per i contrasti /æ/-/ʌ/, /ɜ˞ː/-/ɛ/, /ɑ˞ː/-/ɔ˞ː/ e per il contrasto di controllo /iː/-/ uː/, e A' più bassi rispetto a quelli delle AES ma comunque sufficienti/buoni per i contrasti /iː/-/ɪ/ /ɑː/-/æ/ /ɛ/-/æ/ /uː/-/ʊ/. Il solo contrasto /ʌ/-/ɑː/ ha creato delle difficoltà per le SU che hanno ottenuto un A' al di sotto della soglia minima (Tabella 3).

8.1. Il modello PAM e la capacità discriminativa delle SU

Secondo il modello PAM di Best (1995) dal modo in cui i foni non nativi (fonemi di L2) vengono assimilati ai fonemi nativi dai *naïve listeners* si possono fare delle inferenze su come coppie dei primi possono essere discriminate. Nello specifico, Best (1995) distingue fra foni non nativi che vengono categorizzati con i fonemi nativi (*Categorised*) e foni che non vengono categorizzati con fonemi nativi (*Uncategorised*). Per i foni categorizzati, il PAM prevede tre differenti tipologie di assimilazione alle quali corrispondono tre differenti capacità di discriminazione:

- *Two Category assimilation* (TC), per cui due foni non nativi vengono assimilati fonologicamente e foneticamente a due fonemi nativi differenti chiaramente facili da discriminare.

- *Single Category assimilation* (SC), per cui due foni non nativi vengono assimilati ad un solo fonema nativo sia fonologicamente che foneticamente e vengono entrambi considerati come buoni/poveri esempi delle stesso e, conseguentemente, la loro discriminazione è estremamente difficoltosa.
- *Category Goodness assimilation* (CG), per cui, come nella SC, due foni non nativi vengono assimilati ad un solo fonema nativo sia fonologicamente che foneticamente ma, a differenza di essa, i due foni vengono percepiti come qualitativamente differenti e la discriminazione sarebbe più semplice.

Best (1994, 1995a) riassume nella formula 'TC>CG>SC' i livelli di difficoltà discriminatoria previsti per queste tipologie di assimilazione.

Anche per i foni non categorizzati il PAM prevede tre differenti tipologie di assimilazione con i corrispondenti livelli di discriminazione (Best 1995; Best *et al.*, 2006):

- *Uncategorised-Categorised assimilation* (UC), per cui un fono viene categorizzato in maniera consistente con un fonema nativo mentre l'altro non viene categorizzato in maniera consistente con nessun fonema nativo, portando ad un buon livello di discriminazione poiché si confrontano un fono noto vs. un suono non noto.
- *Uncategorised-Uncategorised assimilation* (UU), per cui entrambi i foni non nativi vengono assimilati a più fonemi nativi e la discriminazione varia da bassa a moderata a seconda che i fonemi nativi a cui i foni non nativi sono stati assimilati siano gli stessi o differiscano.
- *Non-Assimilable* (NA) per cui i due foni non nativi non vengono percepiti come suoni linguistici e vengono discriminati eccellentemente.

In base a queste tipologie di assimilazione, il PAM-L2 (Best & Tyler, 2007) predice che la formazione di nuove categorie fonologiche per i foni di L2 verranno costituite con più probabilità per quei foni che deviano articolatoriamente in maniera maggiore dai fonemi nativi. I test percettivi qui eseguiti sono risultati estremamente funzionali per verificare se il PAM sia applicabile alla tipologia di soggetti da noi studiata, gli SU. In effetti, attraverso l'*identification test* è stato possibile comprendere in che modo i fonemi dell'AE siano stati assimilati ai fonemi dell'IS e tramite l'*oddity discrimination test* è stato possibile verificare se le predizioni del PAM sulla discriminazione dei contrasti individuati siano state realizzate. Nella Tabella 5 si riportano i contrasti individuati tramite l'*identification test*, la corrispondente tipologia di assimilazione predetta dal PAM e i risultati ottenuti nell'*oddity discrimination test*.

Contrasti	PAM	A'
/iː/-/ uː/	TC – eccellente	0.95
/ɑːˠ/-/ɔːˠ/	TC – eccellente	0.94
/iː/-/ɪ/	SC – bassa	0.64
/ɛ/-/ɜːˠ/	CG – intermedia	0.85
/uː/-/ʊ/	CG – intermedia	0.77
/ɑː/-/æ/	UC – buona	0.72
/ɛ/-/æ/	UC – buona	0.67
/ʌ/-/ɑː/	UC –buona	0.42
/æ/-/ʌ/	UU – bassa/buona	0.85

Tabella 5: Tipologie di assimilazione del PAM e A' delle SU

La prima tipologia di assimilazione presa in esame è la TC, in cui rientrano i due contrasti dell'AE /iː/-/uː/ e /ɑːː/-/ɔːː/. Ciascun fonema di L2 dei due contrasti è stato identificato in maniera consistente con un fonema nativo e, come detto in precedenza, questa identificazione è avvenuta sia a livello fonologico che fonetico per cui ci si attenderebbe un'ottima discriminazione da parte delle SU. Effettivamente, in entrambi i casi le previsioni del PAM sono state realizzate in quanto gli A' ottenuti dalle SU sono stati elevati, indicando un'eccellente capacità di discriminazione dei due contrasti (/iː/-/ uː/: A' = 0.95; /ɑːː/-/ɔːː/: A' = 0.94).

Nella seconda tipologia di assimilazione, SC, rientra il contrasto /iː/-/ɪ/, per il quale entrambi i fonemi dell'AE sono stati identificati allo stesso livello fonologico e fonetico con il fonema nativo /i/ con una conseguente discriminazione attesa bassa. La discriminazione del contrasto è risultata essere decisamente inferiore rispetto a quella dei due contrasti precedenti (A' = 0.64), sebbene il risultato indichi una certa sensibilità delle SU alle caratteristiche acustiche delle categorie fonologiche del contrasto, confermando parzialmente le previsioni del PAM.

Nella terza tipologia di assimilazione CG rientrano i contrasti /ɛ/-/ɜːː/ e /uː/-/ʊ/. La discriminazione attesa per questa tipologia è medio/buona in quanto, sebbene i due fonemi di L2 siano assimilati ad uno stesso fonema nativo (rispettivamente /E/ ed /u/), /ɛ/ ed /uː/ si possono considerare un buon esemplare assimilato a livello fonologico e fonetico, ed /ɜːː/, /ʊ/ un esemplare meno buono, ben assimilato da un punto di vista fonologico, ma meno da un punto di vista fonetico. Le SU sono riuscite a discriminare entrambi i contrasti ad un livello buono, come effettivamente previsto da Best.

Relativamente alle tre tipologie di assimilazione sopra descritte, Best (1994, 1995a) afferma che TC, CG, SC sarebbero discriminate nel seguente ordine: 'TC>CG>SC'. Per verificare se la formula 'TC>CG>SC' possa essere applicata ai nostri dati, è stata effettuata una ANOVA univariata con A' come variabile dipendente e con i contrasti di TC, SC e GC a fattore. Poiché i contrasti GC sono due, come detto in precedenza, sono state eseguite due ANOVA univariata: nella prima i contrasti /iː/-/ uː/ (TC) e /iː/-/ɪ/ (SC) sono stati confrontati con /ɛ/-/ɜːː/ (CG), nella seconda con il contrasto /uː/-/ʊ/ (CG). In una prima ANOVA sono stati presi in considerazione i contrasti /iː/-/uː/ (TC), /iː/-/ɪ/ (SC) ed /ɛ/-/ɜːː/ (CG) e la differenza fra i loro A' è risultata essere significativa [$F(2,51) = 19{,}276$ $p < 0{,}01$]. Questa significatività è stata ulteriormente indagata con il test post hoc Tukey, che ha evidenziato che le SU hanno discriminato in maniera significativamente più bassa il contrasto SC rispetto ai contrasti TC e CG ($p < 0{,}01$), e che fra questi ultimi la differenza non è risultata essere significativa ($p > 0{,}01$). Conseguentemente le predizioni di Best sembrano essere state parzialmente confermate: tuttavia in questo caso bisogna tener presente che nel contrasto CG sono contrapposti un fonema vocalico (/ɛ/) ed uno rotacizzato (/ɜːː/) e che questa differenza potrebbe verosimilmente aver favorito la discriminazione dello stesso contrasto.

In una seconda ANOVA i contrasti testati sono stati /iː/-/uː/ (TC), /iː/-/ɪ/ (SC) ed /uː/-/ʊ/ (CG). Anche in questo caso la differenza fra gli A' è risultata significativa [$F(2,51) = 21{,}114$ $p < 0{,}01$], e il post hoc ha rivelato che il contrasto TC è stato discriminato significativamente meglio dei contrasti SC e CG ($p < 0{,}01$) e che fra questi ultimi due la differenza di discriminazione non è significativa sebbene sia molto prossima alla significatività ($p > 0{,}01$ e $p = 0{,}27$). Quindi, anche per quest'analisi si può affermare che la formula 'TC>CG>SC', nel complesso, è stata verificata.

Nella tipologia di assimilazione UC rientrano invece i contrasti /ɛ/-/æ/, /ɑ:/-/æ/ e /ɑ:/-/ʌ/. Per questa tipologia di assimilazione la discriminazione attesa è buona, in quanto si oppongono foni noti (/ɛ/ e /ɑ:/ identificati con /E/ e /a/) a foni non meglio categorizzati (/æ/ e /ʌ/ rispettivamente assimilati ad /E/-/a/ ed /a/-/O/). La discriminazione dei contrasti /ɛ/-/æ/ e /ɑ:/-/æ/ è stata moderata (/ɛ/-/æ/, A' = 0.67; /ɑ:/-/æ/, A' = 0.72), quindi leggermente al di sotto della previsione del PAM, mentre il contrasto /ɑ:/-/ʌ/ non è stato discriminato dalle SU che hanno ottenuto un A' al di sotto della soglia minima (0.42). In quest'ultimo caso, quindi, le previsioni del PAM sono state disattese, ma ciò non costituisce una novità in letteratura. Per questa tipologia di assimilazione, infatti, già Guion *et al.* (2000) avevano riscontrato una discriminazione bassa suggerendo una revisione del modello teorico soprattutto quando i fonemi di L2 sono molto vicini nello spazio acustico, come sembrano essere effettivamente /ɑ:/-/ʌ/ (cfr. Figura 1).

Infine, il contrasto /æ/-/ʌ/ è ascrivibile alla tipologia UU, poiché il fonema /æ/ è stato categorizzato con i fonema nativi /a/-/E/ e il fonema /ʌ/ è stato identificato con i fonemi dell'IS /a/-/O/. Essendo stati assimilati a due coppie di fonemi dell'IS solo parzialmente uguali, condividendo la sola identificazione con il fonema nativo /a/, la discriminazione attesa per questo contrasto è elevata ed effettivamente è ciò che è accaduto alle SU che hanno ottenuto un A' pari a 0.85.

Se riconsideriamo nel complesso questi risultati, sembra possibile affermare che le SU si comportano come *naïve listeners* rispetto ai fonemi dell'AE in quanto sembrano aver seguito i *pattern* di discriminazione previsto dal PAM per i contrasti della L2. Ciò risponde appieno ad uno degli obiettivi di questo lavoro, che era quello di capire se il *framework* del PAM possa essere applicabile anche in ambito FLA, con parlanti adulti aventi un lungo background scolastico della L2. Se in Guion *et al.* (2000) è stato dimostrato che il PAM può essere applicabile anche nei primi stadi di acquisizione della L2 in ambito SLA, i nostri dati ci permettono di ipotizzare che questo modello teorico può essere legittimamente applicato anche a soggetti adulti aventi una lunga conoscenza della L2 sebbene scolastica. Best & Tyler (2007) affermano che soggetti che hanno avuto un'esposizione *limitata* alla L2 in contesti formali, specie con insegnanti non madrelingua con un forte accento di L1, possono annoverarsi nella tipologia dei *naïve listeners*. Le nostre SU hanno avuto un'esposizione alla L2 di circa 10 anni, quindi da un punto di vista meramente cronologico potrebbero non rientrare in tale tipologia ma, ovviamente, non si possono nemmeno ricondurre alla tipologia di *experienced listeners* di cui si occupa Flege. Ci pare quindi coerente concludere che il PAM può essere applicato anche in contesti formali avanzati[8]. In merito alla recente estensione del modello, PAM-L2, ci proponiamo di verificarne le predizioni nei prossimi lavori.

9. DISCUSSIONE

La formazione di una nuova categoria dell'interlingua è prevista da Flege (1995) per quei fonemi non nativi giudicati come nuovi e differenti rispetto ai fonemi nativi contigui. Flege (1988, 1991 e 1995) si rifà al criterio dell'*equivalence classification* secondo il quale le probabilità di formare una nuova categoria fonetica per un parlante nativo di L1 aumenterebbero con l'aumentare della differenza acustica che il parlante stesso percepisce

[8] Cfr. anche Sisinni & Grimaldi (2009) per conclusioni analoghe con studenti universitari salentini che acquisiscono l'inglese britannico come L2.

fra il suono nativo e quello non nativo. Al contrario, quanto più un parlante nativo di L1 percepisce un suono di L2 come simile ad un fonema nativo, tanto più la formazione di una categoria fonetica corrispondente sarà ostacolata.

Sulla stessa linea Best & Tyler (2007) ipotizzano che la formazione di una categoria fonetica è realizzabile quando il parlante riesce a percepire delle differenze acustiche fra i fonemi di L1 e quelli della L2. Per gli autori queste differenze acustiche sarebbero determinate dai differenti gesti articolatori che producono gli eventi acustici, e quando un parlante riesce ad 'individuare' le differenze articolatorie, una categoria fonetica prima e fonologica poi può essere formata per i fonemi della L2.

Rispetto a queste prospettive, possiamo ritornare a riflettere sui dati dell'*identification test*, che offrono una descrizione dettagliata di come i fonemi dell'AE sono stati percepiti rispetto ai fonemi dell'IS (cfr. 7.1): in base ai *fit index*, infatti, si possono fare delle inferenze sui fonemi che sono risultati più facilmente discriminabili. In particolare un *fit index* basso ci dice che i fonemi non nativi sono stati percepiti dalle SU come acusticamente differenti dai fonemi nativi, e verosimilmente presuppongono un processo di facilitazione nella formazione delle categorie fonetiche. Al contrario, i fonemi con un *fit index* elevato sono stati identificati come assolutamente uguali ai fonemi nativi: in questo caso sarà più difficile per gli apprendenti creare nuove categorie fonetiche della L2. Poiché secondo il PAM la nozione di categoria fonetica può essere correlata al livello percettivo ci pare interessante capire la ricaduta che la formazione di una tale categoria può avere anche a livello articolatorio.[9]

Dai risultati del nostro test percettivo emerge che a ricevere *fit index* bassi, inferiori a 2,7, sono i seguenti fonemi dell'AE: /æ/, /ɜ˞ː/, /ʌ/ e /ʊ/. Sembra quindi plausibile pensare che per questi quattro fonemi dell'AE le SU abbiano formato delle categorie fonetiche percettive 'forti'. Fra l'altro i risultati in percezione di questo studio sono in linea con quelli ottenuti in Sisinni & Grimaldi (2009), dove, con la stessa metodologia, è stata testata l'abilità di un gruppo di studenti universitari salentini (scelti con i medesimi parametri del gruppo qui analizzato) di percepire i fonemi, questa volta, del British English (BE), e dove lo stesso gruppo di fonemi ora menzionato è risultato un buon candidato alla formazione di categorie percettive di quella L2.

Benché i valori formantici delle aree di esistenza dei fonemi dell'AE e del BE possano in alcuni casi differire leggermente (cfr. Ladefoged, 2001: 43-45), il fatto che due gruppi di parlanti con lo stesso sistema a 5 vocali, testati in momenti diversi, manifestino gli stessi processi percettivi rispetto all'identificazione dei fonemi della L2 ci induce a ipotizzare che ci troviamo di fronte a dinamiche costanti quando ad interagire si trovano un sistema fonologico a 5 vocali e 3 gradi di apertura (con spazi acustici anteriori, posteriori nonché centrali abbastanza liberi) e un sistema a 5 gradi di apertura con un inventario fonologico molto più ricco che va a riempire la maggior parte dello spazio acustico potenzialmente sfruttabile dai parlanti delle lingue naturali. In Sisinni & Grimaldi (2009), proprio sulla base di questi risultati, e sulla linea degli assunti di Flege e Best, abbiamo infatti ipotizzato che "[…] during the growth of the L2 discrimination process the position occupied by the

[9] Per ulteriori discussioni v. Flege *et al.* (1999), Rauber *et al.* (2005), Bion *et al.* (2006).

L2 contrasts in the vowel spectrum increases salience perception, leading the way in which non-native phonemes are perceptually assimilated to native ones".[10]

Sempre in quest'ottica, riconsideriamo ora i dati dell'*oddity discrimination test* discussi in 6.2-6.3, e in particolare i contrasti della L2 esemplificati in (7) – /æ/-/ʌ/, /ɜ˞ː/-/ɛ/, /ɑ˞ː/-/ɔ˞ː/, /iː/-/ uː/ – che sono stati discriminati dalle SU alla stessa stregua delle parlanti native dell'AE, tenendo presente la Figura 1. A parte il contrasto /iː/-/uː/, che oltre ad essere assimilabile a fonemi dell'IS /i/ ed /u/, presenta una salienza percettiva di per sé rilevante (in quanto composto da fonemi collocati in posizioni periferiche prominenti), per tutti gli altri contrasti anche in questo caso bisogna ipotizzare che la capacità discriminativa sia correlata alla sistematica comparazione acustico-percettiva che, durante il processo di acquisizione, il sistema neuro-uditivo dell'apprendente compie fra le porzioni di spazio acustico in cui sono collocati i fonemi della L1 rispetto a quelli L2. I fonemi in (7) si trovano probabilmente in spazi acustici cruciali, rispetto a quelli occupati dai fonemi della L1, funzionali quindi alla generazione di una salienza percettiva rilevante.

I nostri dati, tuttavia, non ci permettono di fare inferenze specifiche in merito alle peculiarità spettro-temporali e percettive che influiscono nell'agevolare la capacità discriminativa di un particolare set di fonemi rispetto ad altri contrasti presenti nella L2, e d'altro canto i modelli di Flege e Best da questo punto di vista non offrono, allo stato attuale, strumenti per individuare i possibili fattori universali implicati nella individuazione di particolari salienze percettive. Pertanto, come per altri versi sostenuto già da Strange *et al.* (1998: 341-342), vi è la necessità di progettare ricerche che cerchino di cogliere le invarianti dei processi percettivi coinvolti durante la formazione di categorie fonetiche e fonologiche di una L2. I risultati di queste ricerche, come è facile intuire, avrebbero una ricaduta importante sulla metodologia di insegnamento della L2. Inoltre sarebbe interessante in futuro comparare in dettaglio i nostri risultati con quelli ottenuti dal confronto di L1 ed L2 fonologicamente ricche, che presentano quindi spazi acustici ristretti.[11]

Chiarito questo punto, possiamo fare un passo indietro e ritornare sui fonemi dell'AE /æ/, /ɜ˞ː/, /ʌ/ e /ʊ/, che sulla base del *fit index* ottenuto attraverso l'IT rappresenterebbero dei buoni esemplari dell'avvenuto sviluppo di categorie fonetiche della L2, e cercare di capire se tale sviluppo percettivo abbia avuto una ricaduta sistematica a livello acustico-articolatorio.

La categoria fonetica /æ/ è stata articolata in una porzione di spazio acustico intermedia rispetto a quelle dei due fonemi nativi /a/ ed /E/ con cui è stata identificata. Inoltre, in termini di valori formantici medi la /æ/ sembra essere stata prodotta con la stessa anteriorità

[10] In Sisinni & Grimaldi (2009), rispetto al presente studio in cui abbiamo utilizzato la soglia del 70%, si è scelta una soglia pari o superiore al 75% per valutare se un fonema della L2 è stato identificato in maniera consistente con un fonema della L1. Si tenga tuttavia presente che in letteratura tale soglia può coerentemente oscillare fra il 70% e il 80% (cfr. Guion *et al.*, 2000; Tsukada *et al.*, 2005).

[11] Altri studi (Escudero, 2005; Morrison, 2006) che prendono in considerazione come L1 sistemi con 5 vocali rispetto a sistemi L2 più complessi hanno, per scopi specifici delle ricerche, analizzato un set di fonemi abbastanza ridotto, per cui non è possibile fare un confronto adeguato con i nostri dati. D'altro canto i nostri dati concordano in parte con quelli ottenuti da Strange *et al.* (1998: 322), in uno studio cross-linguistico che compara l'acquisizione dei fonemi dell'AE da parte di parlanti giapponesi.

delle AES. Per il fonema /ɜ˞ː/, le SU sono riuscite a creare una categoria fonetica sfruttando una porzione di spazio acustico del tutto vuoto nel loro sistema nativo, ovvero la parte centrale, sebbene non siano riuscite a realizzare valori formantici statisticamente equivalenti rispetto a quelli delle parlanti native. Infine, le categorie fonetiche /ʌ/ e /ʊ/ sono state articolate in maniera nativa in quanto entrambe sono state prodotte con valori formantici medi pari a quelle delle AES. Quindi possiamo concludere che, almeno a livello fonetico, per /ʌ/ ed /ʊ/ sia stato raggiunto un *target* percettivo ed articolatorio ideale, quale presupposto di base per l'acquisizione successiva del livello fonologico.

Se accettiamo l'assunto che la capacità di produzione precede la capacità di percepire i fonemi non nativi, dobbiamo ipotizzare che i nostri soggetti abbiano prima sviluppato l'abilità a produrre le categorie fonetiche /ʌ/ ed /ʊ/ per poi distinguerle percettivamente dai fonemi nativi, o al massimo che i due processi si siano sviluppati in parallelo. Tale ipotesi purtroppo contrasta con il fatto che per le categorie fonetiche /æ/ ed /ɜ˞ː/ il goal acustico articolatorio non è stato realizzato pienamente, come dimostrano i valori formantici in Figura 4. I dati in percezione appena riconsiderati, se comparati con quelli in produzione, ci autorizzano ad assumere che il processo di acquisizione del sistema fonologico di una L2 nei parlanti adulti si basi prioritariamente sulla costruzione di categorie fonetiche percettive derivate dalla parametrizzazione acustica del sistema della L1 rispetto a quello della L2.

Come già evidenziato, il PAM si fonda sugli assunti delle teorie percettive definite *top-down* in cui si ipotizza che oggetto della percezione linguistica siano i gesti articolatori: *articulatory gestures* o *intended phonetic gestures* (Liberman & Mattingly, 1985; Galantucci *et al.* 2006). Secondo questi modelli, e in particolare secondo la *Motor Theory of Speech Perception* di Liberman e Mattingly, l'ascoltatore opera un confronto tra le caratteristiche fisiche del segnale ricevuto e i gesti articolatori necessari per riprodurlo, riconoscendo il segnale attraverso i movimenti necessari a compierlo: la percezione e la produzione del linguaggio condividerebbero, quindi, lo stesso insieme di invarianti, e sarebbero strettamente collegate da un 'modulo' innato. Si presume, perciò, che esista un modulo cerebrale specializzato alla decodifica del segnale acustico prodotto dal linguaggio verbale.[12]

Sulla base dei presupposti teorici del PAM e sulle considerazioni sino ad ora fatte, possiamo assumere che il processo di acquisizione della fonologia della L2, almeno per la tipologia di parlanti qui analizzati, si fondi sulla capacità di individuare progressivamente distanze percettive appropriate fra lo spazio acustico occupato dal sistema della L2 rispetto a quello della L1, tali da generare una salienza percettiva forte: ciò permette lo sviluppo di un processo discriminativo. Da qui si genererà gradualmente l'abilità (a) ad estrarre dal segnale acustico i gesti articolatori che lo generano e che sono i primitivi dei segmenti

[12] Sulla stessa linea si colloca il modello della *Fonologia Articolatoria* di Browman & Goldstein (1992) secondo cui le unità fonologiche fondamentali non sono unità astratte, ma gesti articolatori coordinati fra loro. Tali gesti sono simultaneamente unità d'azione fonetiche – nel senso che ciascun gesto comporta il coordinamento di una serie di muscoli e di articolatori nella costrizione del tratto vocale – ed unità di informazione fonologiche, in quanto potenzialmente distintivi nella creazione di opposizioni fonologiche mediante la presenza/assenza di un gesto particolare o differenze di luogo o grado di una medesima costrizione gestuale. Tuttavia il modello non prevede un modulo innato dedicato al processamento del linguaggio verbale.

fonetici (cfr. Liberman & Whalen, 2000: 188), (b) a proiettare adeguatamente queste rappresentazioni motorie sull'apparato fonatorio nella fase di produzione. Non essendo previsto un livello astratto di rappresentazione, una volta che il parlante è in grado di discriminare un segmento fonetico dovrebbe essere anche capace di produrlo servendosi dei gesti articolatori inequivocabilmente ad esso associati (Liberman & Whalen, 2000: 189). Tuttavia questa visione si accorda in parte con i nostri dati. Benché quattro fonemi dell'AE rappresentino dei buoni esemplari dell'avvenuto sviluppo di categorie fonetiche della L2, solo due, /ʌ/ e /ʊ/, sono stati prodotti coerentemente con i gesti articolatori nativi, mentre i gesti articolatori connessi ai fonemi /æ/ ed /ɜ˞ː/ sono solo parzialmente controllati da questi apprendenti (cfr. Figura 4). Se è vero che una categoria fonetica che viene discriminata produce una percezione immediata e distinta, e che i gesti articolatori connessi sono facilmente recuperabili dal segnale acustico grazie a un modulo neurale specifico per il linguaggio verbale (Liberman & Whalen, 2000: 189), allora i nostri apprendenti avrebbero dovuto chiaramente sviluppare delle capacità articolatorie coerenti per tutte e quattro le categorie fonetiche risultate significativamente discriminate.

La nostra idea, in linea con i modelli *bottom-up*, è che al processo discriminativo e di estrazione delle invarianti articolatorie dal segnale acustico segua un processo neurocognitivo che genera una rappresentazione astratta dei primitivi motori associati a una determinata categoria fonetica: solo quando si è raggiunto quest'ultimo stadio il parlante è in grado di fare le giuste proiezioni motorie sull'apparato fonatorio. Ne consegue che i soggetti analizzati in questo studio sono stati in grado di raggiungere un goal articolatorio adeguato solo di quei fonemi della L2 di cui hanno sviluppato l'appropriata rappresentazione astratta dei tratti articolatori minimi che li identificano in modo oppositivo all'interno del sistema linguistico appreso. Mentre per i fonemi /æ/ ed /ɜ˞ː/ bisogna presupporre che le SU abbiano sviluppato solo una abilità discriminativa forte, che dovrà successivamente essere sostanziata dal processo di rappresentazione astratta dei primitivi articolatori, che porterà in una secondo momento al raggiungimento del goal fonatorio.

La contraddittorietà emersa nei nostri dati rispetto a una interpretazione secondo i modelli percettivi *top-down* trova conforto di riflesso in recenti dati neurofisioligici e di *neuroimaging* che hanno riacceso la discussione proprio intorno alle teorie percettive. Mentre da un lato ci sono evidenze che dimostrano un collegamento fra la corteccia motoria, la corteccia pre-centrale e il controllo dei gesti articolatori specifici per il linguaggio (lingua e labbra, per esempio), dall'altro mancano dati certi che supportino l'idea di un modulo corticale specifico dedicato esclusivamente alla percezione e produzione del sistema fonologico-lessicale, dissociato da altri moduli che controllano i processi motori in generale. Al contrario, pare plausibile che i gesti articolatori coinvolti nel linguaggio verbale siano processati dalle aree cerebrali e dagli stessi circuiti coinvolti anche nella percezione e produzione di gesti motori non linguistici (Fadiga *et al.*, 2002; D'Ausilio *et al.*, 2009; Pulvermüller *et al.*, 2006 e letteratura citata).

Per altro verso un'ampia gamma di osservazioni empiriche di tipo psicolinguistico e neurocognitivo (anche sul versante delle afasie) hanno portato ad ipotizzare che i gesti articolatori del tratto vocale siano costituiti da rappresentazioni motorie astratte derivate direttamente dall'elaborazione corticale del segnale acustico – secondo gli assunti dei modelli percettivi *bottom-up* – e non dal processamento diretto dei gesti prodotti dall'apparato vocale (Hickok & Poeppel, 2007). Ad essere coinvolto in queste operazioni sarebbe un network corticale bilateralizzato, con una forte predominanza a sinistra dove la scissura

perisilviana, che costituisce un'interfaccia sensorimotoria, la parte più anteriore del lobo frontale, che probabilmente coinvolge l'area di Broca, e una parte premotoria dorsale costituirebbero un adeguato network neurale per questo tipo di operazione (cfr. Hickok & Poeppel, 2007: 395). Questi autori sostengono con buoni argomenti che tali rappresentazioni motorie astratte fungono da connessione fra la percezione e la produzione dei suoni delle lingue naturali, e che possono essere identificate con i classici 'tratti distintivi' (cfr. Halle 2001, e per altre evidenze neurocognitive Obleser *et al.*, 2004; Eulitz & Lahiri, 2004).

Dal quadro sinteticamente tracciato emerge che per la ricerca futura nel campo della percezione e produzione del linguaggio, sia per la L1 che per la L2, sarà importante sviluppare ipotesi e modelli teorici che cerchino di trovare una sintesi fra i primitivi motori neurobiologici e i primitivi acustico-articolatori che la linguistica teorica ha posto a fondamento del funzionamento delle lingue naturali. L'obiettivo è quello di arrivare a un modello neurolinguistico unificato ed efficacemente predittivo delle rappresentazioni e delle computazioni che stanno alla base del linguaggio verbale. In quest'ottica ricerche multidisciplinari sui processi di acquisizione della L2 possono dare un apporto sicuramente fondamentale.

10. BIBLIOGRAFIA

Avesani, C., Vayra, M., Best, C. & Bohn, O.S. (2008), Fonologia e acquisizione. In che modo l'esperienza della lingua materna plasma la percezione dei suoni del linguaggio?, in *Processi fonetici e categorie fonologiche nell'acquisizione dell'italiano* (L. Costamagna & G. Marotta, editors), Pacini Editore, 15-41.

Baker, W. Trofimovich, P., Mack, M. & Flege, J.E. (2002), The Effect of perceived phonetic similarity on non-native sound learning by children and adults, in *BUCLD 26: Proceedings of the 26th annual Boston University Conference on Language Development* (A. Do, L. Dominguez & A., Johansen, editors), Somerville, MA: Cascadilla Press.

Best, C.T., Faber, A. & Levitt, A. (1996), Assimilation of Non-Native Vowel Contrasts to the American English Vowel System, *Journal of the Acoustical Society of America*, 99, 2602 (A).

Best, C.T. & Tyler, M.D. (2007), Nonnative and second-language speech perception: Commonalities and complementarities, in *Second Language speech learning: the role of language experience in speech perception and production* (M.J. Munro & O.-S. Bohn, editors), Amsterdam: John Benjamins, 13-34.

Best, C.T. (1995), A direct realist view of cross-language speech perception, in *Speech Perception and linguistic experience: Issues in cross-language research* (W. Strange, editor), Timonium, MD: York Press, 171-204.

Bion, R.A.H., Escudero, P., Rauber, A.S. & Baptista, B.O. (2006), Category formation and the role of spectral quality in the perception and production of English front vowels, in *Proceedings of INTERSPEECH 2006,* Pittsburgh, Pennsylvania, September 18-21, 2006, 1363-1366.

Boersma P. & Weenink D. (2005), *Praat: doing phonetics by computer* (Version 4.3.14) [Computer program]. Retrieved May 26, 2005 [from http://www.praat.org/].

Bohn, O.S. & Flege, J.E. (1991), The Production of New and Similar Vowels by Adult German Learners of English, *Studied in Second Language Acquisition*, 14, 131-158.

Browman, C. P. & Goldstein, L. (1992), Articulatory phonology: An overview, *Phonetica*, 49*,* 155-180.

Collins, B. & Mees, I.M. (2003), *Practical Phonetics and Phonology*, New York: Routledge.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C. & Fadiga, L., The Motor Somatotopy of Speech Perception, *Current Biology 19*, 5, 1-5.

Escudero, P. (2005), *Linguistic Perception and Second Language Acquisition Explaining the attainment of optimal phonological categorization,* Utrecht: LOT.

Eulitz, C. & Lahiri, A. (2004), Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition, *Journal of Cognitive Neurosciences*, 16, 577-583.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002), Speech listening specifically modulates the excitability of tongue muscles: a TMS study, *European Journal of Neurosciences*, 15, 399–402.

Flege, J.E., (1995) Second-Language Speech Learning: Theory, Findings and Problems, in *Speech Perception and linguistic experience: Issues in cross-language research* (W. Strange, editor), Timonium, MD: York Press, 233-273.

Flege, J.E., Munro, M. J. & MacKay, I.R.A. (1995), Factors affecting strength of perceived foreign accent in a second language, *Journal of the Acoustical Society of America*, 97, 3125-3134.

Flege, J.E., Bohn, O.S. & Jang, S. (1997a), Effects of experience on non-native speakers' production and perception of English vowels, *Journal of Phonetics*, 25, 437-470.

Flege, J., Frieda, A. & Nozawa, T. (1997b), Amount of native-language (L1) use affects the pronunciation of an L2, *Journal of Phonetics*, 25, 169-186.

Flege, J.E., MacKay, I.R.A., & Meador, D. (1999), Native Italian Speakers' production and perception of English vowels, *Journal of the Acoustical Society of America*, 106, 2973-2987.

Flege, J.E., Schirru, C. & MacKay, I.R.A. (2003), Interaction between the native and the second language phonetic subsystem, *Speech Communication*, 40, 467- 491.

Flege, J.E. & MacKay, I.R.A. (2004), Perceiving vowels in a second language, *Studies in Second Language Acquisition*, 26, 1-34.

Galantucci, B., Fowler, C.A. & Turvey, M.T. (2006), The motor theory of speech perception reviewed, *Psychonomic Bulletin & Review*, 13, 361–377.

Grassi, C., Sobrero, A. & Telmon, T. (1997), *Fondamenti di dialettologia italiana*, Bari: Laterza.

Grimaldi, M. (2003), *Nuove ricerche sul vocalismo tonico del Salento Meridionale. Analisi acustica a trattamento fonologico dei dati*, Alessandria: Edizioni dell'Orso.

Grimaldi, M. (2009), Acoustic correlates of phonological microvariations. The case of unsuspected highly diversified metaphonetic processes in a small area of Southern Salento (Apulia), in *Romance Languages and Linguistic Theory 2006* (Danièle Torck & W. Leo Wetzels, editors), Amsterdam-Philadelphia: John Benjamins, 89-109.

Grimaldi, M., Sisinni, B., Brattico, E., Invitto, S., Resta, D. & Gili Fivela, B., (in stampa), Correlati comportamentali e neurofisiologici nell'acquisizione di contrasti fonologici della L2, in *Atti del XLII Congresso SLI 'Linguaggio e cervello'*, Scuola Normale Superiore di Pisa, 24-27 settembre 2008.

Guion, S.G., Flege, J.E., Ahahane-Yamada, R. & Pruitt, J.C. (2000), An investigation of current models of second language speech perception: the case of Japanese adults' perception of English consonants, *Journal of the Acoustical Society of America*, 107, 2711-2725.

Halle, M. (2002), *From Memory to Speech*, Berlin: Mouton de Gruyter.

Hansen Edwards, J.G. & Zampini, M. (editors), (2008), *Phonology and Second Language Acquisition*, Amsterdam/Philadelphia: John Benjamins.

Hickok, G. & Poeppel D. (2007), The cortical organization of speech processing, *Nature Review Neuroscience*, 8, 393-402.

Højen, A. & Flege, J.E., (2006), Early learners' discrimination of second-language vowels, *Journal of the Acoustical Society of America*, 119 (5), 3072-3084.

Ladefoged, P. (2001), *Vowels and Consonants. An Introduction to the Sounds of Languages*, Malden, MA: Blackwell Publishing.

Leather, J. (1999), Second Language Speech Research: An Introduction, in *Phonological Issues in Language Learning* (J. Leather, editor), Oxford: Blackwell, 1-56.

Lengeris, A. & Hazan, V. (2007), Cross-language perceptual assimilation and discrimination of Southern British English vowels by Greek and Japanese learners of English, in *Proceedings of the 16th International Congress of Phonetic Sciences,* Saarbrücken, Germany, August 6-10, 2007, 1641-1644.

Lenneberg, E.H. (1967), *Biological foundations of language*, New York: John Wiley.

Liberman, A.M., & Mattingly, I.G. (1985), The motor theory of speech perception revised, *Cognition*, 21, 1-36.

Liberman, A.M. & Whalen, D.H. (2000), On the relation of speech to language, *Trends in Cognitive Sciences*, 4, 187-196.

Llisterri, J. (1995), Relationships between Speech Production and Speech Perception in a Second Language, in *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, Sweden. 14-19 August 1995, 92-99.

MacKay, I.R.A, Meador, D. & Flege, J.E. (2001), The Identification of English Consonants by Native Speakers of Italian, *Phonetica*, 58, 103-125.

Mora, J.C. & Fullana, N. (2007), Production and perception of English /iː/-/ɪ/ and /æ/-/ʌ/ in a formal setting: Investigating the effects of experience and starting age, *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, August 6-10, 2007, 1613-1616.

Morrison, G.S. (2006), *L1 & L2 Production and Perception of English and Spanish Vowels. A Statistical Modelling Approach*, PhD Thesis, University of Alberta.

Obleser, J., Lahiri, A. & Eulitz, C. (2004), Magnetic brain response mirrors extraction of phonological features from spoken vowels, *Journal of Cognitive Neuroscience* 16, 31–39.

Piske, T., MacKay, I.R.A. & Flege, J.E. (2001), Factors affecting degree of foreign accent in an L2: a review, *Journal of Phonetics*, 29, 191-215.

Piske, T., Flege, J.E., MacKay, I.R.A. & Meador, D. (2002), The Production of English Vowels by Fluent Early and Late Italian-English Bilinguals, *Phonetica*, 59, 49-71.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O. & Shtyrov, Y. (2006), Motor cortex maps articulatory features of speech sounds, in *Proceedings of National Academy of Sciences, USA*, 103, 7865–7870.

Rauber, A.S., Escudero, P., Bion, R.A.H. & Baptista, B.O. (2005), The Interrelation between the Perception and Production of English Vowels by Native Speakers of Brazilian Portuguese, in *Proceedings of INTERSPEECH 2005*, Lisbon, Portugal, September 4-8, 2005.

Sisinni, B. & Grimaldi, M. (2009), Second language discrimination vowel contrasts by adults speakers with a five vowel system, in *10th Annual Conference of the International Speech Communication Association (ISCA), Interspeech*, Brighton, September 6-10, 2009, 1679-1682.

Strange, W. (2007), Cross-language phonetic similarity of vowels: Theoretical and methodological issues, in *Language experience in second language speech learning: In honor of James Emil Flege* (O.-S. Bohn & M.J. Munro, editors), Amsterdam-Philadelphia: Benjamins, 35-55.

Strange, W., Yamada, R.A., Kubo, R., Trent, S.A., Nishi, K. & Jenkins, J.J. (1998), Perceptual Assimilation of American English vowels by Japanese listeners, *Journal of Phonetics*, 26, 311-344.

Tsukada, K., Birdsong., D., Bialystok, E., Mack, M., Sung, H. & Flege, J.E. (2005), A developmental study of English vowel production and perception by native Korean adults and children, *Journal of Phonetics*, 33, 263-290.

Tsukada, K. (2006), Cross-language perception of word-final stops in Thai and English, *Bilingualism: Language and Cognition*, 9, 309-318.

LA DIMENSIONE TEMPORALE IN TRE TIPI DI PARLATO: UN CONFRONTO TRA ARABO E ITALIANO

Dalia Gamal I. Abou El Enin
Università di Ain Shams (Il Cairo)
daliagamal60@hotmail.com

1. SOMMARIO

Questo contributo si basa su materiale di varia natura in cui viene osservata l'organizzazione temporale a livello macroprosodico. Il corpus consiste di brani di parlato letto e di parlato semispontaneo elicitato tramite il metodo *Map Task* in arabo e italiano L1 e in italiano L2. Per un maggiore approfondimento vengono analizzati due brevi brani di parlato spontaneo di interviste televisive in arabo e italiano L1. Ciascun tipo di parlato analizzato in questo lavoro presenta caratteristiche diafasiche particolari che si possono intravvedere nel ritmo. Dunque abbiamo l'ambizione di condurre una prima analisi dell'impiego della dimensione temporale in italiano L2 da parte di apprendenti guidati di provenienza egiziana e di fare un confronto con i dati che emergono dal materiale prodotto nella loro lingua prima e nella lingua bersaglio.

Nella scelta del parametro durata come unico parametro studiato in questa ricerca si è consapevoli del fatto che indici prosodici, quali gli accenti intonativi e la posizione delle sillabe toniche nell'unità tonale, sono tra i fattori che incidono sull'organizzazione temporale del discorso connesso in italiano. È noto anche che le alternanze di durata di sillabe toniche e atone determinano la ritmicità del parlato (Bertinetto & Magno Caldognetto, 1993; Nespor, 1994; Giordano, 2006). Ma il presente contributo mira a focalizzare su fenomeni macroprosodici delineati dal parametro durata, soprattutto la durata delle sequenze articolate e la durata e la distribuzione delle pause.

Lo sguardo rimane fisso sull'ipotesi che la lingua prima abbia un'influenza percepibile sulla prosodia in L2.

2. MATERIALE E METODO

Il materiale qui analizzato è di varia natura diatopica e diafasica. Si divide in tre tipi: parlato letto che si compone di un brano in arabo standard di 115 sillabe fonologiche e di tre brani in italiano di 412 sillabe (260+75+77); parlato (semi)spontaneo selezionato da tre dialoghi *Map Task* in italiano L1, italiano L2 e arabo cairota L1; due brani di parlato spontaneo registrato dalla televisione, uno in arabo egiziano, di 205 sillabe, e l'altro in italiano (169 sillabe), ciascuno della durata di 35 secondi circa; infine è stato anche chiesto ai nostri 5 soggetti di recitare alcuni versetti del Corano per poter formare una idea più chiara sulla loro cultura ritmica nella lingua madre, in quanto la buona recitazione del Corano profila una struttura temporale particolare che riassumeremo a suo tempo (§ 3.1.2.2.).

Tutti gli informatori egiziani, di cui verrà esposto il profilo linguistico-culturale, hanno letto i brani del giornale e del telegiornale arabo e italiano. Per il confronto con i nativi sono state analizzate le notizie 'originali' prodotte da *speaker* professioniste.

I due brani estratti dal telegiornale italiano riguardano in realtà la stessa notizia, leggermente modificata nel testo e presentata in due edizioni diverse del tg, di cui una è per i non

udenti (notizia 'a'). Quest'ultima versione si presenta all'orecchio con un ritmo molto lento rispetto all'altra edizione 'b'.

I testi per la lettura sono stati dati agli informatori qualche minuto prima della registrazione e gli è stato permesso di darci uno sguardo veloce. Il compito del *Map Task*, invece, è stato svolto secondo le regole conosciute della tecnica di elicitazione, secondo le quali gli interlocutori arrivano a scoprire i nomi delle icone sulle mappe e la natura del compito man mano durante l'interazione verbale, per cui l'elemento della sorpresa la rende il più possibile naturale.

Come campione del parlato spontaneo si è optato per le interviste televisive in L1 a causa della difficoltà di incoraggiare gli informatori a parlare la L2 con un compaesano senza che ci sia un compito ben definito; in tal caso la comunicazione spontanea in L2 rappresenterebbe per i parlanti una situazione assai fittizia. Il caso tipico in cui il parlante non nativo dovrebbe parlare la lingua straniera con la massima naturalezza possibile è il caso della comunicazione interetnica, che resta in fin dei conti poco equilibrata e spesso si registrano casi in cui il parlante straniero si appoggia sulle scelte lessicali dell'interlocutore nativo (Wiberg, 2004), il che suscita dubbi sul suo grado di naturalezza, anche se resta il modello più comune, nella vita reale, della comunicazione spontanea in lingua straniera.

Gli informatori italiani sono due ricercatrici all'università, di età tra i 35 e i 45 anni, la prima (**F Ca**) è della Calabria e la seconda (**F Na**) è della Campania. Non è stato possibile controllare ulteriormente la variazione diatopica del corpus in italiano L1, in quanto i parlanti italiani, incluso il locutore del materiale spontaneo, sono di origine varia, ma queste due informatrici presentano un livello culturale molto simile e sono entrambe di provenienza meridionale.

Gli apprendenti guidati registrati per questo studio sono cinque, quattro informatrici e un solo informatore; tutti lavorano nel Dipartimento di italiano nella Facoltà di lingue e sono tutti provenienti dalla capitale, il Cairo:

- **FA**: 28 anni, assistente universitaria, conosce l'italiano da 11 anni, ha passato 8 mesi in Italia, un anno e mezzo circa prima del momento della registrazione.
- **FB**: 32 anni, dottore di ricerca in lingua italiana (studiata per 15 anni), ottimo inglese e diversi soggiorni in Italia, complessivamente un anno.
- **MC**: 29 anni, ha passato 5 anni in Italia per conseguire il dottorato ed era tornato alla distanza di un anno dalla registrazione.
- **FD**: 31 anni, soggiorno di qualche mese in Italia, studia la lingua italiana dall'età di 17 anni.
- **FE**: 24 anni, assistente, esperienza dell'italiano di 7 anni.

Il materiale elicitato con *Map-Task* è stato registrato quando l'informatrice FA aveva 21 anni circa e quindi aveva un'esperienza della lingua italiana di appena quattro anni.

3. DATI E DISCUSSIONE

Il materiale fonico è stato diviso in foni con l'ausilio del programma Wavesurfer ed ai fini del presente studio sono state misurate le durate delle sillabe, delle catene foniche e delle pause piene (non silenti) e vuote (silenti). A differenza della 'catena fonica', la quale racchiude la fonazione tra due pause vuote, la 'sequenza articolata' è la porzione che non contiene neanche le pause piene (Pettorino & Giannini, 2005). Sono state calcolate la velocità di articolazione in ogni sequenza (= n. sillabe / loro durata) e la proporzione delle sequenze articolate rispetto alle pause all'interno di ogni brano o turno dialogico.

3.1 Lo stile letto

3.1.1 Italiano letto

Per la lettura del brano del giornale e di 2 notizie del tg esponiamo prima i grafici delle parlanti native poi quelli dei cinque apprendenti egiziani.

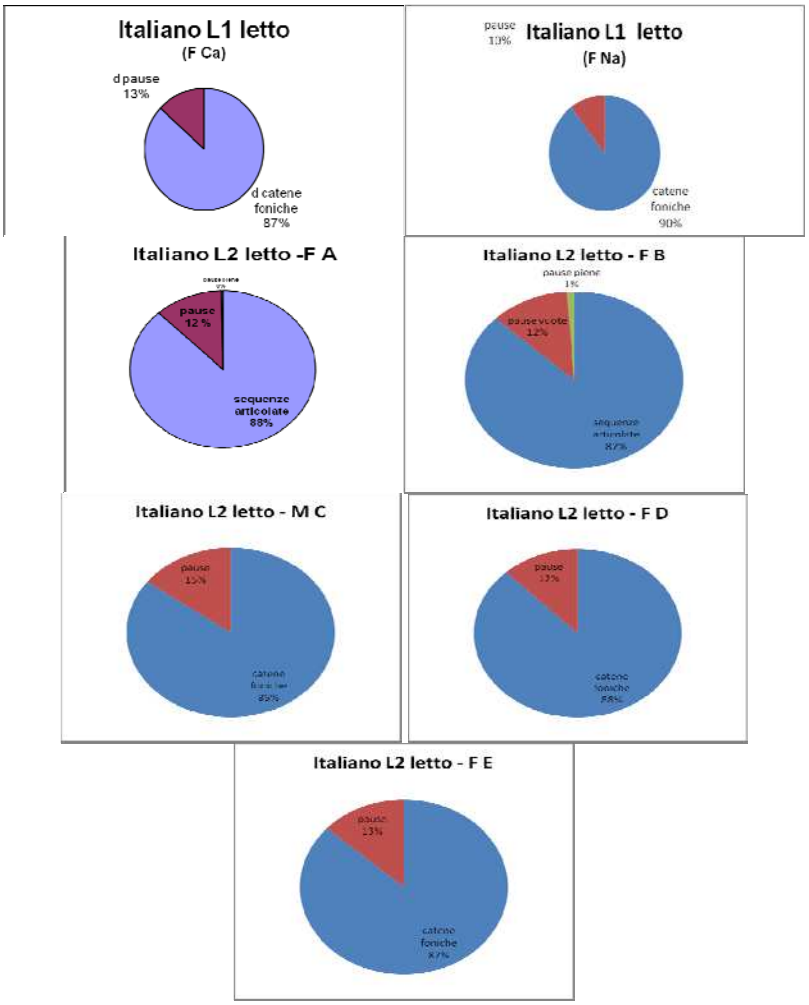


Figura 1: percentuale delle pause vuote rispetto alle catene parlate
in italiano L1 e L2

I grafici degli apprendenti raggruppano tutto il materiale letto in italiano, in quanto il brano del giornale e i due testi del telegiornale non presentano grandi differenze di occorrenza delle pause (5% in FA e M C, 3% in F B e F E e 0% in F D); inoltre nessuno dei due tipi di testo presenta sistematicamente attraverso le produzioni di tutti gli informatori una maggior percentuale delle pause. La velocità di articolazione media nelle produzioni degli informatori sono, rispettivamente, 7 e 6 sill/sec in L1 e 6, 8, 6, 5 e 5.6 sill/sec in L2.

Si rileva che la percentuale delle pause in L2 non è molto diversa dalla loro percentuale di occorrenza in italiano L1. Infatti, i parlanti non esercitano in questa situazione lo sforzo che richiede il parlato spontaneo per la selezione lessicale e la strutturazione degli enunciati; e questo si evince dalla distribuzione delle pause all'interno della stringa (vedi appendice per la trascrizione ortografica completa dei sette informatori). Si nota che in generale la scansione delle parole da parte degli informatori egiziani è molto simile a quella di F Na, che ha una pausazione più ricca rispetto a F Ca. Quest'ultima, tuttavia, colloca le sue poche pause in posizioni scelte anche da F Na e gli altri soggetti. Si veda la pausazione di FA, che a parte i numerali e l'abbreviazione che l'informatrice probabilmente non conosce, presenta una collocazione quasi identica a quella di F Na. In sette posizioni troviamo un pieno accordo tra i sette informatori nell'introdurre una pausa; sono posizioni segnalate da segni d'interpunzione (il punto a fine frase, i due punti, le parentesi).

Inoltre, si nota la difficoltà di alcune parole che bloccano gli apprendenti, soprattutto la parola 'stratosferici'. Essendo a un livello avanzato, gli apprendenti sanno avvalersi della punteggiatura e delle congiunzioni per riprendere fiato e non interrompono un sintagma se non in casi estremi, davanti alle parole nuove o inaspettate. In merito, è stato rilevato un dato interessante, che, tuttavia, non si presenta così frequentemente da poterlo considerare un fenomeno. Infatti, si è osservato che, a differenza dei nativi, gli apprendenti si bloccano all'interno del sintagma e persino in mezzo alla parola. Come esempio di interruzione del primo tipo abbiamo il sintagma 'mercato ipotecario americano', durante la lettura del quale gli egiziani si bloccano direttamente prima della parola nuova, 'ipotecario', mentre F Na si ferma prima dell'intero sintagma aggettivale; inoltre, F Ca a un certo punto interrompe la lettura (davanti a una percentuale) e fa una lunga pausa per tornare a leggere la frase dall'inizio {audio 1}. L'unico esempio del secondo tipo di interruzione si rileva nella produzione dell'informatrice FD che divide la parola 'cinquanta' dopo la prima sillaba e poi continua la lettura senza tornare indietro:

Es.: l'uno ***<p>*** *virgola cinq+* ***<p>*** *virgola cinq****<p>*** *uantanove percento.* {audio 2}

Un altro calcolo temporale che evidenzia differenze o somiglianze tra L1 e L2 riguarda la durata media delle pause e delle sequenze articolate (tabella 1), che molto spesso coincidono con catene foniche per l'assenza quasi totale delle pause piene in questo stile.

	Catene foniche	Pause
F Ca	2593	486
F Na	1889	271
F A	1412	228
F B	1765	259
M C	1576	319
F D	1801	287
F E	1933	350

Tabella 1: durata media (ms) delle catene foniche e delle pause nello stile letto in arabo e italiano L2

Anche qui le differenze nella lunghezza delle catene foniche e delle pause non si rive-

lano grandi. Solo F Ca, che si fermava poco, presenta catene molto lunghe e di conseguenza pause di respiro assai estese. Le tabelle 2 e 3 ci danno invece la durata complessiva del brano giornalistico e delle notizie, rispettivamente.

F Ca	49.883
F Na	51.857
F A	1:00.298
F B	58.255
M C	52.019
F D	1:05.054
F E	1:01.725

Tabella 2: durata assoluta (min:sec.ms) della lettura del brano del giornale italiano (informatrici italiane e informatori egiziani)

	a	b
Professionista	17.962	11.728
F A	14.626	15.690
F B	15.898	15.785
M C	13.852	14.155
F D	17.164	17.089
F E	15.343	14.816

Tabella 3: durata (sec) delle notizie a e b del tg italiano

La produzione della notizia 'a' occupa più tempo da parte della speaker professionista perché è estratta da una edizione speciale per i non udenti. Solo il dato assoluto della durata complessiva del brano letto segnala una generale differenza tra nativi e non. FD impiega una durata totale vicina alla durata della versione lenta della professionista (**ItPr**). Nel primo brano ciò è dovuto non solo alla maggiore durata sillabica, ma anche alle pause frequenti da entrambe le parlanti; nel brano b, invece, lo scarto tra le medie sillabiche è il dato determinante: 160 ms per ItPr e 194 per FD.

3.1.2 Arabo letto

Questa parte del corpus ci serve a questo punto per sapere quali abitudini ritmiche i nostri apprendenti riprendano, eventualmente, dalla loro lingua prima nella lettura della lingua seconda. Ciò si può cercare nella scansione delle parole in gruppi di respiro. Si potrà anche fare un confronto tra la velocità d'articolazione e l'estensione relativa delle pause all'interno di ogni unità di testo nelle due lingue.

Il materiale in arabo letto si divide in un brano del telegiornale e in 3 versetti del Corano. L'arabo scritto è normalmente l'arabo standard e non quello dialettale. Si osserva infatti che a livelli medi e alti di scolarizzazione l'arabo strettamente dialettale risulta molto difficile alla lettura, dato che non si è soliti usarlo nei contesti scritti.

Inoltre, la recitazione del Corano non si presenta normalmente come quella degli altri testi scritti, perché oltre al valore e la stima di cui gode il testo, ci sono delle regole ben precise che regolano la lunghezza dei singoli foni nei vari contesti. Quindi, la dimensione

temporale in tale tipo di recitazione è molto importante e il lettore deve essere attento ad aumentare o ridurre le durate in relazione al contesto segmentale successivo (vedi *infra* § 3.1.2.2.).

3.1.2.1 Tg arabo

Nel materiale di arabo letto la durata totale delle pause oscilla tra l'8% e il 18% nelle produzioni dei cinque informatori; Le velocità di articolazione in arabo, sono, rispettivamente, 6, 7, 7.5, 7 e 8 sill/sec (figura 2). Rispetto alla speaker professionista (figura 4) le informatrici F A e F B presentano le occorrenze delle pause più simili.

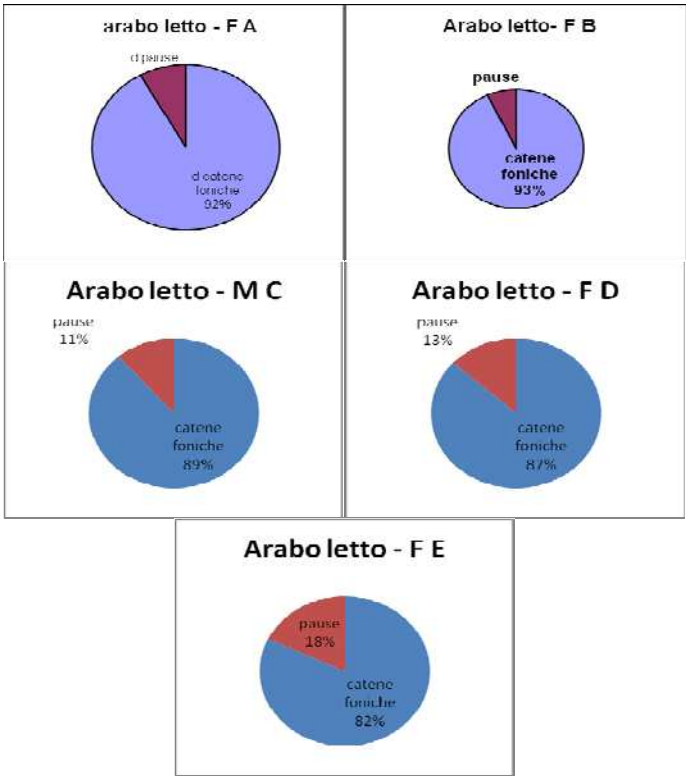


Figura 2: la distribuzione del tempo tra pause e catene foniche in arabo L1

	Catene foniche	Pause
F A	2475	342
F B	3384	308
M C	2684	370
F D	2349	396
F E	2220	534

Tabella 4: durata media (ms) di catene foniche e pause

Le medie assolute delle catene foniche non presentano notevoli scarti tranne per FB. Riguardo alla posizione delle pause in questa parte del corpus gli informatori fanno pause prima delle congiunzioni (affinché, e), tra il nome e l'apposizione e tra il verbo e il soggetto se quest'ultimo è costituito da un sintagma lungo. Il numero delle pause in ogni produzione non è basso (in media 8 pause) e si registrano alcune disfluenze.

Dopo la presentazione dei dati relativi alla recitazione coranica possiamo avanzare riflessioni generali sul materiale letto.

3.1.2.2. Recitazione del Corano

Secondo le regole di buona recitazione le vocali dovrebbero subire un allungamento uguale al doppio della durata di una vocale lunga in ogni posizione in cui viene seguita dalla consonante occlusiva glottidale; inoltre, la nasale, che per effetto di coarticolazione subisce il cambiamento del punto d'articolazione, viene allungata quanto una vocale lunga, quindi, secondo gli esperti, si raddoppia di durata. Tali regole sono state sempre seguite sin dall'inizio della rivelazione coranica nel settimo secolo, ma i linguisti hanno cominciato a scriverle in maniera sistematica nei loro trattati circa un secolo dopo (Hilāl, 1996).

Prima delle registrazioni si è saputo che gli informatori MC e FE non sanno recitare il Corano secondo tali regole; di conseguenza si vedrà che la loro velocità di articolazione è più alta rispetto agli altri (vedi tabella 5). I tre versetti includono 42 sillabe.

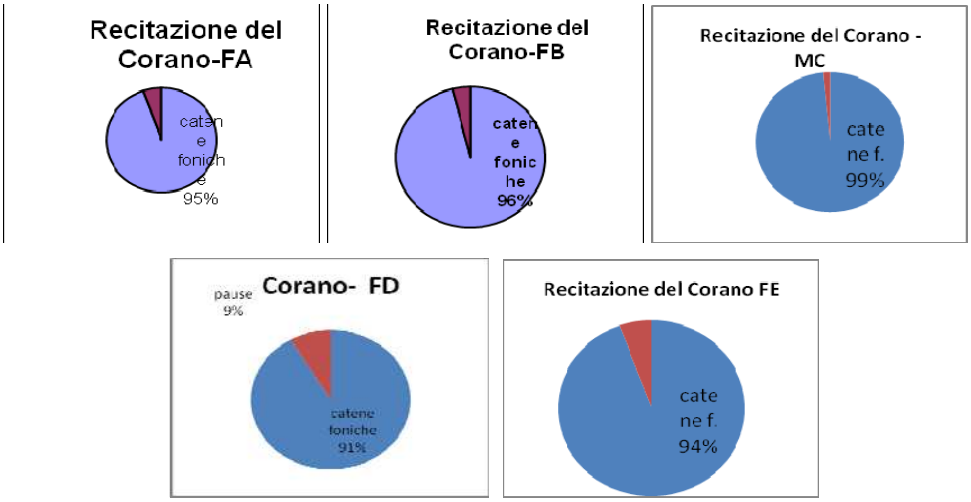


Figura 3: percentuale delle pause nella recitazione del Corano

Si osserva che in questo tipo particolare di recitazione i nostri informatori dedicano una porzione molto ridotta alle pause, malgrado la lunghezza dei silenzi non faccia parte delle regole di recitazione. Ora vediamo la seguente tabella delle velocità di articolazione:

	L1		L2
	tg	Corano	
F A	6	2	6
F B	7	4	8
M C	7.5	7	6
F D	7	4	5
F E	8	6	6.5

Tabella 5: velocità di articolazione (sillabe per secondo)
degli apprendenti nelle due lingue

Prima di tutto spiccano i valori notevolmente bassi relativi alla recitazione del Corano da parte degli apprendenti che rispettano le regole ritmiche di questo tipo di produzione orale. Tale calcolo è il risultato dei frequenti prolungamenti dei foni. Siamo, dunque, di fronte a 3 parlanti con uno sfondo ritmico variegato nella L1 e che presentano una consapevole manipolazione della durata a seconda del contesto.

Rispetto alla L2, però, il comportamento dei cinque informatori non si delinea omogeneo, nel senso che non si può affermare che nella lingua straniera, a livelli avanzati di apprendimento, la produzione orale è più lenta o, nel miglior dei casi, uguale alla velocità nella L1. Si osserva persino che l'informatrice FB presenta una velocità più grande nella L2. Integrando tale dato alle durate medie delle sue catene foniche e pause, si evidenzia una produzione con catene di media lunghezza e pause non di alta percentuale né di lunga durata rispetto agli altri informatori.

Si osservi, per contro, che la velocità d'articolazione di MC non è la più alta. Si ricorda che egli ebbe la maggiore permanenza in Italia di tutti gli informatori, permanenza per scopi di approfondimento della lingua italiana.

3.1.3 Arabo e italiano del telegiornale

Prima di passare al secondo tipo di parlato occorre spendere qualche parola sulla produzione professionale dei tg, nei limiti di quanto ci offra il nostro materiale.

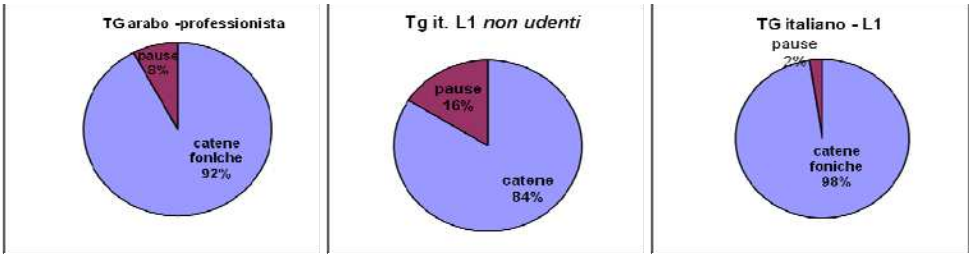


Figura 4: percentuale delle pause nei telegiornali delle speaker professioniste.

La velocità d'articolazione in arabo è 6,7 sill/sec e in italiano 6 sill/sec nella lettura normale e 5 sill/sec nell'edizione per i non udenti 'a'.

Si nota al primo sguardo quanto si giochi sulle pause per rendere la lettura più lenta e intelligibile.[1] La pausa qui è l'indice prosodico che esercita un ruolo maggiore rispetto alla velocità di elocuzione. ItPr impiega 14952 ms per proferire 161 foni nella versione speciale 'a'; nel brano 'b' i 168 foni occupano 12315ms.

	arabo	italiano 'b'	italiano 'a'
catene foniche	3159	6158	831
pause	265	119	171

Tabella 6: durata media (ms) di catene foniche e pause
nei tg delle professioniste

3.2 Lo stile elicitato

Di questo tipo di parlato abbiamo analizzato turni dialogici di registrazioni *Map Task* in arabo (16 turni) e in italiano L2 (16 turni) dell'informatrice FA, e anche 11 turni del dialogo A01 del corpus AVIP-API, registrato in area campana.

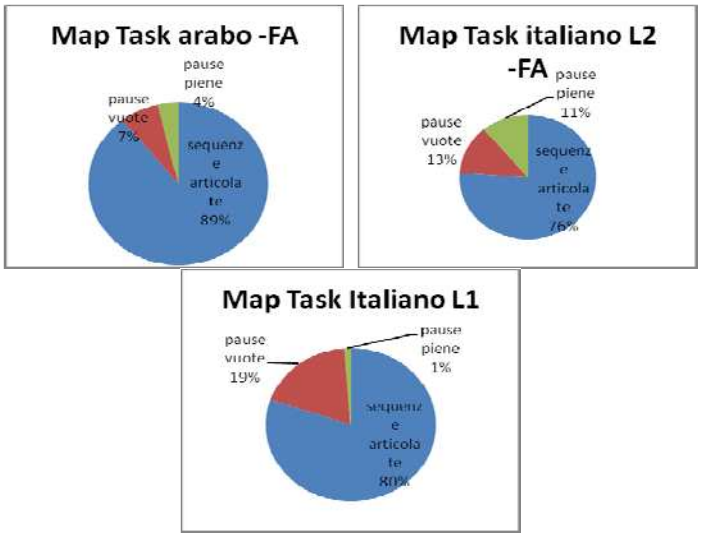


Figura 5: pause e sequenze articolate nel Map Task in arabo e italiano

La velocità d'articolazione media di FA non presenta grandi differenze in L1 e L2 (in arabo è 6,5 e in italiano 6,3 sill/sec). Il parlante napoletano ha una velocità media di 6,4 sill/sec. Invece, la maggiore differenza tra le tre produzioni si segnala nella percentuale delle pause all'interno delle catene foniche e la diversa occorrenza delle pause piene e vuote.

[1] Infatti, per esperienza diretta nell'insegnamento si osserva che gli studenti di lingua italiana nel corso di laurea riescono a capire meglio le notizie di questo tipo, invece l'ascolto dell'edizione 'normale' li porta alla frustrazione.

Map Task		
	media d sequenza	media d pausa
italiano L1	1801	724
italiano L2	960	377
arabo L1	1577	236,2

Tabella 7: medie delle sequenze articolate e delle pause nel compito della mappa

È interessante in questa tabella la lunghezza delle pause del parlante nativo, ma l'ascolto della registrazione lascia capire che il *giver* si fermava a lungo per aspettare la partner della conversazione. In questo stile vincolato a un compito guidato le pause aumentano molto rispetto allo stile letto, ma la velocità di articolazione dell'informatrice FA (6,3 sill/sec). si avvicina ai valori esibiti nello stile letto (6 sill/sec). Ciò si potrebbe ricondurre alla natura del compito, poiché l'informatrice aspettava a lungo risposte e *feedback* da parte del *follower*. Le pause si riscontrano anche prima di alcuni nomi delle icone o quando il *follower* chiede una ulteriore spiegazione. Questi due casi, in realtà, limitano la libertà dell'informatrice che nel primo caso viene sorpresa dalla richiesta del partner e deve ripetere la parola in modo più chiaro, molto spesso inconsapevole della differenza tra le due mappe; e nel secondo caso deve sforzarsi per cercare un altro modo di spiegare.

3.3 Il tipo spontaneo

Per le analisi del tipo spontaneo abbiamo due brani di interviste in italiano e in arabo L1. Si tratta di due parlanti maschi di età attorno ai cinquanta anni. L'intervista in arabo ad un parlante egiziano è estratta da una trasmissione di approfondimento giornalistico della *all-news* araba *Al Jazeera*, mentre il brano in italiano è preso da una breve intervista in *Rainews 24*.

La complessa dinamica del parlato si presenta qui in primo piano ed esercita un influsso grande sulla pausazione dei due parlanti, soprattutto quello egiziano. Questi sembra attento a dire il più possibile nel tempo limitato concessogli. Trapela dalla manipolazione della f_0 e dell'intensità (I) e dalla distribuzione delle pause quanto lui si aspetti una prossima interruzione da parte della conduttrice, la quale ogni tanto dà la parola a un altro ospite che, va ricordato, esprime sempre l'opinione opposta. Perciò il parlante si trova palesemente sotto stress.

Infatti, si osserva nel brano in arabo che cinque delle sette pause di respiro intercorrono all'interno della proposizione, interrompendo la struttura sintattica.

(1) *si sono resi conto <p respiro> della sua difficoltà*
(si veda la traduzione di tutto il brano in appendice)

In questo esempio la proposizione viene spezzata tra verbo e complemento; in altri casi viene interrotta tra soggetto e predicato. Inoltre, il nostro parlante egiziano, invece di fermarsi alla fine delle proposizioni compiute, continua il gruppo di respiro con una nuova struttura che egli interrompe di nuovo per riprendere fiato. Nell'esempio seguente le sillabe in grassetto portano movimenti melodici di salita (vedi *infra* la figura 6).

(2) *e si sono accontentati […] <p respiro> di un'altra visione. Ha**mas** è in disaccordo con questa visione. In **ques**to contesto si verifica la polemica tra l'Egitto e Hamas. Perciò <p respiro> non è accettabile […]*

Tale scansione della struttura sintattica non è diffusa in arabo. In un lavoro precedente (Gamal, 2006), in un corpus di parlato semispontaneo in lingua araba, è stata rilevata l'integrazione degli indici di scansione prosodica, quali l'allungamento prepausale e la riprogrammazione di f_0 e I, con la pausa nella divisione della catena parlata in unità tonali, le quali nello stesso tempo corrispondono a unità dal senso compiuto.

In questo materiale televisivo, però, il tentativo di scansione in unità prosodiche comporta maggiori difficoltà a causa del conflitto tra la distribuzione delle pause e la collocazione dei fenomeni melodici rilevanti (cioè livelli di f_0 marcati o grandi scarti nei movimenti). Se facciamo una rappresentazione grafica dei valori di f_0 e I che accompagnano le parole iniziali e finali in ogni proposizione principale ci rendiamo conto che i valori alti di f_0 e i movimenti melodici di salita vengono sfruttati per attirare l'attenzione alle nuove strutture che nello stesso tempo trasmettono idee nuove.

Nella figura 6 si nota che l'andamento melodico ed energetico è molto funzionale alla scansione sintattica. Se ci atteniamo a questi due parametri e escludiamo le pause nella divisione in unità tonali, troviamo che tali unità dal senso compiuto presentano, esclusa la seconda, un livello melodico iniziale alto e un andamento finale discendente che riflette la conclusione dell'idea.

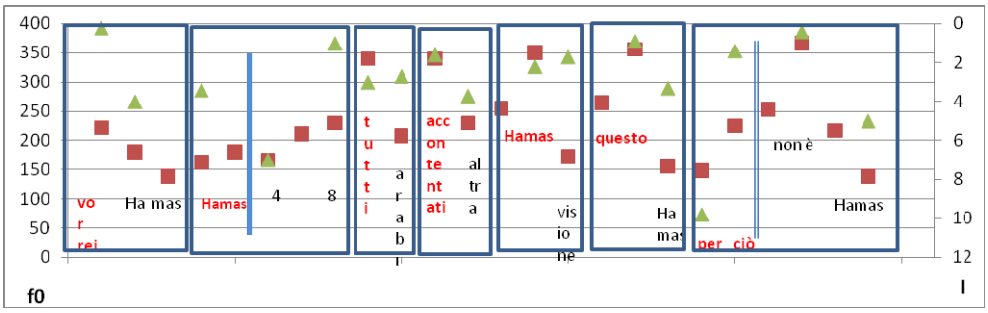


Figura 6: valori di f0 (rettangoli rossi) e di I (triangoli verdi)
sulle parole iniziali (in grassetto rosso) e parole finali di proposizione
(i riquadri blu dividono delineano i confini sintattici)

Le due linee verticali nel grafico indicano due pause che incidono sull'andamento tonale. Ciascuna delle due pause si colloca dopo una salita di sospensione. Tale salita, infatti, serve a collegare l'elemento che precede l'interruzione agli elementi successivi e in fin dei conti serve a ridurre l'effetto dell'interruzione e a mantenere la parola, facendo aspettare gli altri locutori. Dunque, il parlante riesce a servirsi dei mezzi prosodici per poter portare avanti la comunicazione, manipolando alla meglio energia e melodia. Cerchiamo, dunque, di capire perché le pause non si collochino in posizioni che segnalino confini sintattici. Le pause, in realtà, presentano ad un primo ascolto una distribuzione casuale che di conseguenza risulterebbe poco funzionale come indice che possa guidare l'ascoltatore nella comprensione del messaggio linguistico. Invece, come segnalato riguardo alle due pause indicate nella figura 6, si osserva che le cinque pause che interrompono la struttura sintattica, si collocano dopo una salita di f_0 o dopo una tenuta melodica. Quindi, la pausa che si colloca nella posizione meno aspettata, preceduta da una sospensione melodica,

rende l'ascoltatore in continua attesa e, mentre ai confini sintattici i partner dell'intervista dovrebbero credere di poter riprendere la parola, il parlante continua senza sosta e rimanda la ripresa del fiato ai punti in cui è difficile interromperlo. Si nota anche che alla fine di questo intervento analizzato i valori di f_0 profilano una discesa fino ai minimi del *range* del parlante. Solo in questo momento la conduttrice riprende la parola.

Da tutto ciò si può concludere che anche qui la distribuzione delle pause, osservata insieme all'andamento melodico, è funzionale a mantenere la parola in questa situazione comunicativa poco rilassata. L'argomento, naturalmente, richiede ulteriori analisi di materiale più grande.

Nell'intervista italiana la coincidenza con confini sintattici non è sistematica; soltanto due delle cinque pause vuote ricadono a fine proposizione. Tuttavia, il dato più interessante per il confronto con l'intervista araba riguarda l'occorrenza delle pause vuote, le quali sono più frequenti in italiano rispetto all'arabo. Ciò rivela che il parlante italiano, che si presenta come l'unico ospite nella breve intervista, riesce a prendersi tempo per pensare, riprogrammare il proprio discorso e cercare le parole adatte senza la preoccupazione di perdere la parola. Nella figura 7 si vede la percentuale delle pause nelle due interviste.

La frequenza delle pause nel brano in italiano, tuttavia, contrasta con quanto si prevede in un contesto in cui il parlante conosce bene l'argomento (Magno Caldognetto & Vagges, 1992) e sembra ben preparato a dare una informazione precisa ai telespettatori (si veda la trascrizione ortografica in appendice).

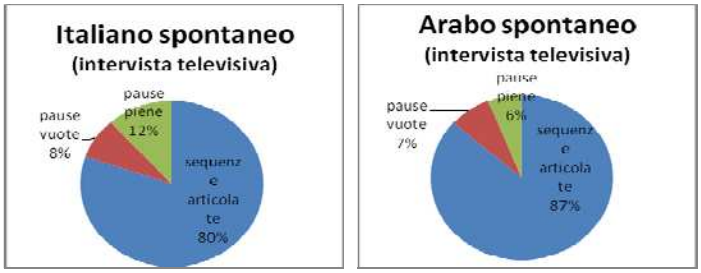


Figura 7: pause piene, pause vuote e sequenze articolate nei campioni di parlato spontaneo in arabo e italiano L1

	catene foniche	pause
arabo	2410	304
italiano	2501	305

Tabella 8: durata media (ms) delle catene foniche e delle pause nel materiale televisivo spontaneo arabo e italiano.

Si nota nella tabella 8 che le sequenze articolate sono molto simili nella lunghezza. Inoltre, le durate medie delle pause sono identiche nei campioni televisivi nelle due lingue, anche se i grafici nella figura 7 mostrano che i due tipi di pause sono di occorrenza diversa. Tuttavia, i grafici confermano la solita occorrenza delle pause piene nel discorso spontaneo a differenza del parlato letto. La velocità di articolazione in italiano è 6 sill/sec e in arabo 6,7 sill/sec.

4. CONCLUSIONI

In letteratura i dati sul ritmo presentano un quadro assai variegato e ben lontano dalla conclusione. Anche lo stile letto continua a costituire una miniera per le indagini fonetiche presentando dati che necessitano di interpretazioni profonde e di ulteriori analisi nelle varie lingue del mondo (Mohd Don *et al.*, 2008).

In questo contributo è stato analizzato un corpus che all'inizio della ricerca sembrava potesse dare alcune risposte, anche parziali, al quesito sull'impiego del tempo in L2 in rapporto alla lingua di partenza e possibilmente anche relativamente alla lingua d'arrivo. Ci si è trovati, tuttavia, davanti a una serie di dati che portano alla riflessione sulla forte sensibilità degli indici del ritmo alla variazione situazionale.

In effetti, i cinque parlanti egiziani ci rivelano ancora una volta quanto sia difficile cercare regolarità generalizzabili in fatti di ritmo (Ramus, 2002). Anche nello studio della L2 ci si trova costretti ad osservare la produzione linguistica come un prodotto originale che non è condizionato esclusivamente dalle capacità e le conoscenze linguistiche dell'apprendente, ma anche dalla natura comunicativa della produzione linguistica.

È intuitivo che l'apprendente eserciti maggiore sforzo per la progettazione del parlato non letto, ma è altrettanto vero e evidente nei dati che si sente sotto la lente d'ingrandimento, anche nel parlato letto, e quindi in quest'ultimo tipo di parlato impiega maggior tempo rispetto ai nativi per poter procedere con il minimo rischio di fare errori. Sotto questo aspetto si può considerare che la lettura sia un tipo di situazione comunicativa a 'senso unico', non solo perché trasmette informazioni, ma, nel caso dell'apprendente, anche perché rivela capacità linguistiche.

Nella parte palesemente confrontabile del corpus, il tipo letto, i sette informatori non professionisti, ciascuno nella sua categoria di parlante nativo o di apprendente, evidenziano delle differenze nei valori di durata medi e assoluti. Ma nello spazio temporale diverso occupato da ciascuno di loro per finire il proprio compito comunicativo la distribuzione della fonazione e dei silenzi è generalmente corrispondente. Inoltre, la collocazione delle pause è molto funzionale alla trasmissione del senso e al superamento delle difficoltà lessicali.

Dunque la dimensione temporale in L2, a un livello avanzato di conoscenza linguistica, resta più estesa, ma manipolata e organizzata in modo armonico e funzionale in modo da dare un quadro di dati, relativi, simili a quelli dei parlanti della L1.

Ma se cerchiamo di interpretare i nostri dati esclusivamente alla luce del confronto con la lingua prima torneremo, non di rado, con le mani vuote. La stessa lingua prima, e l'esperienza linguistica in generale offrono un patrimonio ritmico-situazionale che invece di condizionare e trapelare negativamente nella L2, rende l'apprendente più sensibile alle esigenze della situazione in tale lingua straniera.

In futuro, una analisi delle pause in un confronto tra apprendenti di livelli vari potrebbe spiegare meglio lo sviluppo del controllo e l'impiego della pausazione. Nel parlato spontaneo dei media si può analizzare il ritmo di vari parlanti in comunicazioni più lunghe che presentino reazioni e strategie comunicative diverse.

5. BIBLIOGRAFIA

Bertinetto, P.M. & Magno Caldognetto, E. (1993), Ritmo e intonazione, in *Introduzione all'italiano contemporaneo. Le strutture* (A. Sobrero, editor), Roma-Bari: Laterza, 141-192.

Gamal, D. (2006), La prosodia direttiva in italiano L2. Studio pilota, in *Analisi prosodica. Teorie, modelli e sistemi di annotazione*, Atti del 2° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Salerno, 30 novembre – 2 dicembre 2005 (R. Savy & C. Crocco, editors), Torriana (RN): EDK Editore, 189-202.

Giannini, A. & Pettorino, M. (1999), I cambiamenti dell'italiano radiofonico negli ultimi 50 anni: aspetti ritmo-prosodici e segmentali, in *Atti delle 9^e^ Giornate del G.F.S. – Venezia, 1998* (R. Delmonte & A. Bristot, editors), Venezia: Università di Ca' Foscari, 65-81.

Giordano, R. (2006), Note sulla fonetica del ritmo dell'italiano, in *Analisi prosodica. Teorie, modelli e sistemi di annotazione*, Atti del 2° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Salerno, 30 novembre – 2 dicembre 2005 (R. Savy & C. Crocco, editors), Torriana (RN): EDK Editore, 233-244.

Hilāl, A. H. (1996), أصوات اللاغه العربية [F*onologia della lingua araba*], Cairo: Wahba.

Magno Caldognetto, E. & Vagges, K. (1992), Le pause quali indici diagnostici per lo stile del parlato spontaneo, in *Atti delle 2^e^ Giornate di Studio del G.F.S.- Arcavacata, 1991* (J. Trumper & L. Romito, editors), Roma: Esagrafica, 97-106.

Mohd Don, Z., Knowles, G. & Yong, J. (2008), How Words can be Misleading: A Study of Syllable Timing and 'Stress' in Malay, *The Linguistics Journal*, vol. 3; sul sito: http://www.linguistics-journal.com/August_2008_zmd.php.

Nespor, M. (1994), *Fonologia*, Bologna: Il Mulino.

Pettorino, M. & Giannini, A. (2005), Analisi delle disfluenze e del ritmo di un dialogo romano, in *Italiano parlato: analisi di un dialogo* (F. Albano Leoni & R. Giordano, editors), Napoli: Liguori, 89-104.

Ramus, F. (2002), Acoustic Correlates of Linguistic Rhythm: Perspectives, in *Proceedings of the International Conference Speech Prosody 2002* (B. Bel & I. Marlien, editors), Aix-en-Provence, France, 115-120.

Wiberg, E. (2004), Strategie interazionali dell'apprendente nel dialogo tra nativo e non-nativo, in, *Atti del Convegno Nazionale sul Parlato Italiano – Napoli, 2003* (F. Albano Leoni, F. Cutugno, M. Pettorino & R. Savy, a cura di), Napoli: D'Auria, CD-ROM, articolo G11.

APPENDICE

Le pause **<p>** che non superano la solita pausa di respiro del parlante non sono considerate lunghe e se superano tale durata vengono indicate come pause lunghe <pl>. <p respiro> è la pausa riempita da un chiaro rumore di inspirazione. Il segno + segue le disfluenze.

FA

"Fannie Mae **<p>** e Freddie Mac: **<p>** sembra il titolo di un musical di Broadway **<p>** ed è invece la nuova minaccia **<p>** che incombe **<p>** sui mercati finanziari **<p>** e sulle prospettive economiche globali. **<p respiro>** Le cifre sono terribili: **<p>** le due istituzioni hanno dimensioni enormi **<p>** (complessivamente, 5.300 milioni di dollari, **<p respiro>** il trenta+ il trentotto percento del Pil **<p>** Usa) **<p>** e detengono **<p>** o garantiscono circa la metà **<p>** dell'intero mercato **<p>** ipotecario americano. **<p>** Esse però **<p>** sono intrinsecamente fragilissime, **<p>** perché **<p>** grazie alla garanzia statale **<p>** hanno potuto spingere il loro **<p>** indebitamento **<p respiro>** a livelli strato<laringalizzazione>+ stratosferici: **<p respiro>** una stima di **<p>** Ubs indica per Freddie **<p>** un rapporto tra patrimonio e totale **<p respiro>** dell'attivo di appena **<p>** <eeh>l'uno virgola **<p>** cinquantanove percento a fine dicembre. **<p>** Livelli **<p>** da brividi.

F B:

"Fannie Mae e Freddie Mac: **<p>** sembra il titolo di un musical di Broadway **<p>** ed è invece la nuova minaccia che incombe sui mercati finanziari e sulle s+ **<p>** prospettive economiche globali. **<p>** Le cifre **<p>** sono terribili: **<p>** le due istituzioni hanno dimensioni enormi **<p>** (complessivamente, **<pl>** <eeh> 5.300 milioni di dollari, **<p>** il trentotto percento del **<p>** Pil Usa) e detengono o garantiscono circa la metà dell'intero mercato **<p>** ipotecario americano. **<p>**. Esse però sono intrinsecamente fragilissime, **<p>** perché grazie **<p>** alla garanzia statale hanno potuto spingere il loro indebitamento **<p>** a livelli **<p>** stratosferici: **<p>** una stima di **<p>** Ubs indica per Freddie un rapporto tra patrimonio e totale dell'attivo di appena **<p>** l'uno<oo> virgola cinquantanove percento a fine dicembre. **<p>** Livelli da brividi.

M C

"Fannie Mae e Freddie Mac: **<p>** sembra il titolo di un musical **<p>** di musical di Broadway **<p>** ed è invece la nuova minaccia che incombe sui mercati finanziari **<p>** e sulle prospettive economiche globali. **<p>** Le cifre sono terribili: **<p>** le due istituzioni **<p>** hanno dimensioni enormi **<p>** (complessivamente, **<p>** 5.300 milioni di dollari, **<p>** il trecento per+ il trentotto percento del Pil Usa) **<p>** e detengono o garantiscono **<p>** circa la metà dell'intero mercato ipotecario americano. **<p>** Esse però sono intrinsecamente fragilissime, **<p>** perché grazie alla garanzia statale **<p>** hanno potuto spingere il loro indebitamento a livelli stratosferici: **<p>** una stima di Ubs indica per Freddie un rapporto **<p>** tra patrimonio e totale **<p>** e totale dell'attivo di appena **<p>** l'uno virgola cinquantanove percento **<p>** a fine dicembre. **<p>** Livelli da brividi.

FD

"Fannie Mae e freddie Mac: **<p>** sembra il titolo di un musical di Broadway **<p>** ed è invece la nuova minaccia **<p>** che incombe sui mercati finanziari **<p>** e sulle prospettive economiche globali. **<p>**

Le cifre sono terribili: **<p>** le due istituzioni hanno dimensioni enorme **<p>**

(complessivamente, **<p>** 5.300 milioni di dollari, **<p>** il 38**<p>** % **<p>** del Pil Usa) **<p>** e detengono **<p>** o garantiscono circa la metà dell'intero mercato **<p>** ipotecario americano. **<p>** Esse però **<p>** sono intrinsecamente fragilissime, **<p>** perché grazie alla garanzia statale **<p>** hanno potuto spingere il loro indebitamento a livelli **<p>** stratosc'+**<p>** stratosferici: **<p>** una stima di Ubs **<p>** indica per Freddie un rapporto tra patrimonio **<p>** e totale dell'attivo di appena l'uno **<p>** virgola cinq+ **<p>** virgola cinq**<p>** uantanove percento a fine di+ **<p>** di+ **<p>** a fine dicembre. **<p>** Livelli da brividi.

FE
"Fannie Mae **<p>** e freddie Mac: **<p>** sembra il titolo di un musical di Broadway **<p>** ed è invece la nuova minaccia **<p>** che incombe sui mercati finanziari **<p>** e sulle prospettive economiche globali. **<p>**
Le cifre sono terribili: **<p>** le due istituzioni hanno dimensioni enormi **<p>** (complessivamente, **<p>** 5.300 milioni di dollari, **<p>** il trento+ **<p>** il trentotto percento del Pil **<p>** Usa) **<p>** e detengono **<p>** o garantiscono circa la metà dell'intero mercato **<p>** ipoteca+ **<p>** ipotecario **<p>** americano. **<p>** Esse però sono intrinsecamente fragilissime, **<p>** perché grazie alla garanzia statale **<p>** hanno potuto spingere il loro indebitamento **<p>** a livelli **<p>** stratosferici: **<p>** una stima di Ubs **<p>** indica per Freddie un rapporto tra patrimonio **<p>** e totale dell'a+ **<p>** dell'attivo di appena **<p>** l'uno **<p>** virgola cinquantanove percento a fine dicembre. **<p>** Livelli da brividi.

F Ca
"sembra il titolo di un musical di Br**oa**dway ed è invece la nuova minaccia che incombe sui mercati finanz***ia****ri* e sulle prospettive economiche globali. **<p respiro>** Le cifre sono terribili: **<p>** le due istituzioni hanno dimensioni en**o**rme **<p>** (complessivamente, **<p>** 5.300 milioni di dollari, **<p>** il trentotto percento del Pil Usa) **<p>** e detengono o garantiscono circa la metà dell'intero mercato ipotecario americano. **<p respiro>** Esse però sono intrinsecamente fragil*issime*, perché grazie alla garanzia statale hanno potuto spingere il loro indebitamento a livelli stratosferici: **<p respiro>** una stima di Ubs indica per Freddie un rapporto tra patrimonio e totale a+ dell'attivo di appena+ **<pl>** una stima Ubs indica per Freddie un rapporto tra patrimonio e totale dell'attivo di appena l'uno v+ ci+ virgola cinquantanove percento **<p respiro>** a fine dicembre. **<p>** Livelli da brividi.

F Na
"Fannie Mae **<p>** e Freddie Mac: **<p>** sembra il titolo di un musical di Broadway **<p>** ed è invece la nuova minaccia che incombe sui mercati finanziari **<p>** e sulle prospettive economiche globali. **<p>** Le cifre sono terribili: **<p>** le due istituzioni hanno<oo> **<p>** dimensioni enormi **<p>** (complessivamente, 5.300 milioni di dollari, **<p respiro>** il trentotto percento del Pil **<p>** <eeh> Usa) **<p>** e detengono **<p>** o garantiscono circa la metà dell'intero **<p>** <eeh> mercato ipotecario americano. **<p>** Esse però sono intrinsecamente fragilissime, **<p>** perché grazie alla garanzia statale **<p>** hanno potuto spingere il loro indebitamento a livelli stratosferici: **<p>** una stima di Ubs **<p>** indica per Freddie **<p>** un rapporto tra patrimonio **<p>** e totale dell'attivo di appena **<p respiro>** l'uno virgola cinquantanove percento a fine dicembre. **<p>** Livelli da brividi.

Dal telegiornale italiano:
parlante professionista
(Non udenti):
Sono<oo> scattati **<p>** oggi **<p>** i saldi nelle grandi città **<p>** italiane. **<p>** È **<p>** il primo banco di prova **<p>** del duemila e nove **<p>** su fiducia **<p>** e acquisti **<p>** degli italiani. **<p>** Nel<ee> **<p>** periodo **<p>** natalizio, **<p>** dice **<p>** Confcommercio, **<p>** i **<p>** consumi **<p>** hanno **<p>** tenuto.

E andiamo in Italia ora. Al via oggi i saldi nelle grandi città, il primo banco di prova del 2009 su fiducia **<p respiro>** e consumi dopo che nel periodo natalizio, dice Confcommercio, c'è stata una sostanziale tenuta.

F A
Sono scattati oggi i saldi nelle grandi città italiane. **<p respiro>** È il primo banco di prova del duemila e nove su fiducia e acquisti **<p>** degli italiani. **<p>** Nel periodo natalizio, **<p>** dice Conf+ Confcommercio, **<p>** i consumi hanno tenuto.

E andiamo in Italia ora. **<p>** Al via **<p>** oggi i saldi nelle grandi città, **<p>** il primo banco di prova del 2009 su fiducia e consumi **<p respiro>** dopo che nel periodo natalizio, **<p>** dice Confcommercio, **<p>** c'è stata una sostanziale tenuta.

F B
Sono scattati oggi i saldi nelle grandi città italiane. **<p>** È il primo banco di prova **<p>** del duemila e nove **<p>** su fiducia e acquisti degli italiani. **<p>** Nel periodo **<p>** natalizio, **<p>** dice Con+ Confcommercio, **<p>** i consumi hanno tenuto.

E andiamo in Italia ora. **<p>** Al via oggi **<p>** i saldi nelle grandi città, **<p>** il primo banco di prova **<p>** del 2009 **<p>** su fiducia e consumi **<p>** dopo che nel periodo **<p>** natalizio, **<p>** dice Confcommercio, **<p>** c'è stata una sostanziale tenuta.

M C
Sono scattati oggi i saldi nelle grandi città italiane. **<p>** È il primo banco di prova del duemila e nove **<p>** su fiducia e acquisti degli italiani. **<p>** Nel periodo natalizio, **<p>** dice Conf+ **<p>** Confcommercio, **<p>** i consumi hanno tenuto.

E andiamo in Italia ora. **<p>** Al via oggi i saldi nelle grandi città, **<p>** il primo banco di prova del 2009 **<p>** su fiducia e consumi **<p>** dopo che nel+ **<p>** nel periodo natalizio, **<p>** dice Conf+ **<p>** Confcommercio, **<p>** c'è stata una sostanziale tenuta.

F D
Sono scattati oggi i saldi nelle grandi città italiane. **<p>** È il primo banco di prova del duemila **<p>** e nove **<p>** su fiducia e acquisti degli italiani. **<p>** Nel periodo **<p>** natalizio, **<p>** dice Conf**<p>**commercio, **<p>** i consumi hanno tenuto.

E andiamo in Italia ora. **<p>** Al via oggi i saldi nelle grandi città, il primo banco di prova del 2009 su fiducia e consumi **<p>** dopo che nel periodo natalizio, dice Confcommercio, **<p>** c'è stata una sostanziale tenuta.

F E
Sono scattati oggi **<p>** i saldi nelle grandi città italiane. **<p>** È il primo banco di prova del duemila **<p>** e nove **<p>** su fiducia e acquisti degli a+ italiani. **<p>** Nel periodo natalizio, **<p>** dice Confcommercio, **<p>** i consumi hanno tenuto.

E andiamo in Italia ora. **<p>** Al via oggi i saldi nelle grandi città, **<p>** il primo banco di prova del 2009 su fiducia e consumi dopo che nel periodo natalizio, **<p>** dice Confcommercio, **<p>** c'è stata una sostanziale tenuta.

Testo del brano tratto dall'intervista televisiva italiana:
Mah <p> sul problema pochi giorni fa c'è stata una riunione a livello comunale con gli esponenti interessati. <p respiro> <eeh> è stato sottolineato<oo> <eeh> la gravità del problema e il<ll>l'importo eccessivo che viene speso per un servizio che non è <p respiro> <eeh> calzante rispetto alle esigenze dell'utenza. <p respiro> Si è deciso di dare determinate soluzioni come quella di integrare <p respiro> il discorso di taxi <eh> del noleggio con conducente e incentivare le cooperative di tassisti affinché <p> <eh> diano un servizio esclusivamente ai disabili.

Arabo letto:

A) Traduzione del brano:

Dalla sua parte il sovrano saudita, Abdullah Ibn Abdel-Aziz, ha convocato un vertice straordinario dei capi di stato dei paesi del Golfo a Riyad oggi per esaminare l'offensiva israeliana contro Gaza e i mezzi atti a far raggiungere una tregua umanitaria per far entrare gli aiuti alimentari e medicinali nella Striscia di Gaza. Una fonte del Ministero degli Esteri saudita ha riferito che i leader dei Paesi del Consiglio di Cooperazione del Golfo hanno accolto l'invito del re saudita.

B) Trascrizione in arabo del brano letto da ciascun informatore:

Speaker professionista

من جهته دعا خادم الحرمين الشريفين العاهل السعودى الملك عبد الله بن عبد العزيز**<p>** إلى قمة خليجية طارئة**<p>** فى الرياض**<p>** اليوم**<p>** لبحث العدوان الإسرائيلى على غزة**<p respiro>** و السبل الكفيلة للتوصل إلى هدنة إنسانية**< p respiro>** لدخول المساعدات الغذائية و الطبية إلى القطاع. **<p respiro>** و أفاد مصدر فى وزارة الخارجية السعودية بأن قادة دول مجلس التعاون **<p>** استجابوا لدعوة خادم الحرمين الشريفين**<p>** بعقد القمة الخليجية الطارئة

FA

من جهته دعا خادم الحرمين الشريفين العاهل السعودى الملك عبد الله بن عبد العزيز**<p respiro>** إلى قمة خليجية طارئة فى الرياض **<p>** ال**<p>** فى الرياض اليوم**<p>** لبحث العدوان الإسرائيلى على غزة**<p>** و السبل الكفيلة للتوصل إلى هدنة إنسانية لدخول المساعدات الغذائية و الطبية إلى القطاع. و أفاد مصدر فى وزارة الخارجية السعودية**<p>** بأن قادة دول مجلس التعاون **<p>** استجابوا لدعوة خادم الحرمين الشريفين**<p>** بعقد القمة الخليجية الطارئة

FB

من جهته دعا خادم الحرمين الشريفين العاهل السعودى الملك عبد الله بن عبد العزيز**<p>** إلى قمة خليجية طارئة فى الرياض اليوم**<p>** لبحث العدوان الإسرائيلى على غزة**<p>** و السبل الكفيلة للتوصل إلى هدنة إنسانية لدخول المساعدات الغذائية و الطبية إلى القطاع. **<p>** إلى القطاع. **<p>** و أفاد مصدر فى وزارة الخارجية السعودية بأن قادة دول مجلس التعاون **<p>** استجابوا لدعوة خادم الحرمين الشريفين**<p>** بعقد القمة الخليجية الطارئة

MC

من جهته دعا خادم الحرمين الشريفين **<p>** العاهل السعودى الملك عبد الله بن عبد العزيز**<p>** إلى قمة خليجية طارئة فى الرياض اليوم**<p>** لبحث العدوان الإسرائيلى على غزة**<p>** و السبل**<p>** الكفيلة للتوصل إلى هدنة إنسانية لدخول المساعدات الغذائية و الطبية إلى القطاع. **<p>** و أفاد مصدر فى وزارة الخارجية السعودية بأن قادة دول مجلس التعاون**<p>** استجابوا لدعوةخا+ خادم الحرمين الشريفين**<p>** بعقد القمة الخليجية الطارئة

FD

من جهته**<p>** دعا خادم الحرمين الشريفين العاهل السعودى الملك عبد الله بن عبد العزيز**<p>** إلى قمة خليجية طارئة فى الرياض اليوم لبحث العدوان الإسرائيلى على غزة**<p>** و السبل الكفيلة للتوصل إلى هدنة إنسانية**<p>** لدخول المساعدات الغذائية و الطبية إلى القطاع. **<p>** و أفاد مصدر فى وزارة الخارجية السعودية**<p>** بأن قادة دول مجلس التعاون استجابوا لدعوة خادم الحرمين الشريفين**<p>** بعقد القمة الخليجية الطارئة

FE

من جهته دعا الخادم الحرمين الشريفين**<p>** العاهل السعودى الملك عبد الله بن عبد العزيز**<p>** إلى قمة خليجية طارئة فى الرياض اليوم**<p>** لبحث العدوان الإسرائيلى على غزة**<p>** و السبل الكفيلة للتوصل إلى هدنة إنسانية لد +لدخول المساعدات الغذائية و الطبية إلى القطاع. **<p>** و أفاد مصدر فى وزارة الخارجية السعودية**<p>** بأن قادة الدول**<p>** مجلس**<p>** بأن قادة الدول و مجلس التعاون **<p>** استجابوا لدعوة خادم الحرمين الشريفين بعقد قم+ **<p>** القمة الخليجية الطارئة

Testo del brano tratto dall'intervista televisiva araba:

A) Traduzione in italiano:

<aah> una altra volta vo+/ vorrei mettere questo nel contesto del disaccordo strategico tra la visione egiziana e la visione di Hamas. La visione di Hamas <p respiro> <eh> la liberazione di tutto il suolo nazionale dalla terra al fiume, '48, una visione legittima, ma tutti gli arabi l'hanno provata prima <p respiro> e si sono resi conto <p respiro> della sua difficoltà e/ e il costo grande <p respiro> che <eeh> dovranno pagare+/ che <eeh> dovranno pagare il popolo palestinese e i popoli arabi. E si sono accontentati, tutti gli arabi, tra cui i radicali arabi <p respiro>, di un'altra visione. Hamas è in disaccordo con questa visione. In questo contesto si verifica la polemica tra l'Egitto e Hamas. Perciò <p respiro> non è dunque accettabile che si dica <p respiro> che l'Egitto debba adottare il punto di vista di Hamas.

B) Testo arabo:

<أه> مرة أخرى أ+/ أريد أن أضع هذا فى سياق الخلاف الاستيراتيجى ما بين الرؤية المصرية و رؤية حماس. رؤية حماس < p respiro> <أه> تحرير كامل التراب الفلسطينى من الأرض إلى النهر 48. رؤية مشروعة لكن كل العرب جربوها قبل ذلك < p respiro> و تبينوا < p respiro> صعوبتها و و التكلفة الكبيرة < p respiro> التى <أه> تعود/ التى <أه> تعود على الشعب الفلسطينى و على الشعوب العربية بسببها. و ارتضى العرب بما فيهم راديكاليي العرب < p respiro> رؤية أخرى. حماس تختلف مع هذه الرؤية. فى هذا السياق يحدث الجدل ما بين مصر و حماس و من ثم < p respiro> ليس من المقبول إذن أن يقال < p respiro> أن تتبنى مصر وجهة نظر حماس.

TECNOLOGIE
DEL PARLATO

ALCUNE CONSIDERAZIONI SULL'IMPORTANZA DEGLI ASPETTI DINAMICI NELLA PERCEZIONE, PRODUZIONE ED ELABORAZIONE DEL PARLATO

Piero Cosi
Istituto di Scienze e Tecnologie della Cognizione - Sede di Padova 'Fonetica e Dialettologia'
Consiglio Nazionale delle Ricerche
e-mail: piero.cosi@pd.istc.cnr.it

1. SOMMARIO

In questo lavoro vengono sinteticamente illustrati alcuni dei più significativi apporti tecnologici che nel corso degli ultimi anni sono stati influenzati dalla dimensione temporale del parlato nel campo dell'analisi del segnale vocale, della sintesi della voce da testo scritto e del riconoscimento automatico del segnale verbale. Per quanto riguarda la realizzazione di facce parlanti animate, sono discussi poi alcuni esempi dell'influenza degli aspetti dinamici nella percezione e nella'interpretazione delle espressioni facciali e più in generale degli intenti comunicativi, nella trasmissione di emozioni, stati d'animo e atteggiamenti, nell'interazione faccia a faccia.

2. INTRODUZIONE

La dimensione temporale è un elemento costitutivo non solo dei meccanismi di produzione del parlato, intervenendo, a livello segmentale, nella determinazione delle durate e nella pianificazione e nel controllo di tutti i gesti articolatori e, a livello soprasegmentale, nell'allineamento dei contorni intonativi con le parti dell'enunciato, ma anche, nella percezione del segnale verbale e, più in generale, nell'interpretazione di un qualsiasi atto comunicativo. Ad esempio, sia la configurazione delle caratteristiche facciali che la sincronizzazione delle azioni facciali sono importanti nell'espressione e nel riconoscimento delle emozioni (Cohn, 2007).

In questa breve rassegna si fa ampio riferimento a due lavori precedentemente presentati. In particolare per quanto riguarda il parlato si fa riferimento a *50 Years of Progress in Speech and Speaker Recognition Research*, presentato da Sadaoki Furui nel 2005 e pubblicato in *ECTI Transactions on Computer and Information Technology* (Furui, 2005) e, per quanto riguarda la percezione delle espressioni facciali e più in generale degli intenti comunicativi, si fa riferimento a *Foundations of human-centered computing: Facial expression and emotion*, presentato da Jordan F. Cohn nel 2007 e pubblicato in *Proceedings of the International Joint Conference on Artificial Intelligence* (Cohn, 2007).

3. DIMENSIONE TEMPORALE E SPEECH TECHNOLOGY

Per quanto riguarda il Trattamento Automatico del Linguaggio (*TAL*), e in particolare, l'analisi del segnale vocale, la sintesi della voce da testo scritto e il riconoscimento automatico del segnale verbale, gli apporti tecnologici più significativi che nel corso degli ultimi anni sono stati influenzati dalla dimensione temporale del parlato sono illustrati schematicamente in Figura 1, dove sono evidenziate le caratteristiche specifiche che nel corso degli anni hanno reso sempre più affidabili queste tecnologie.

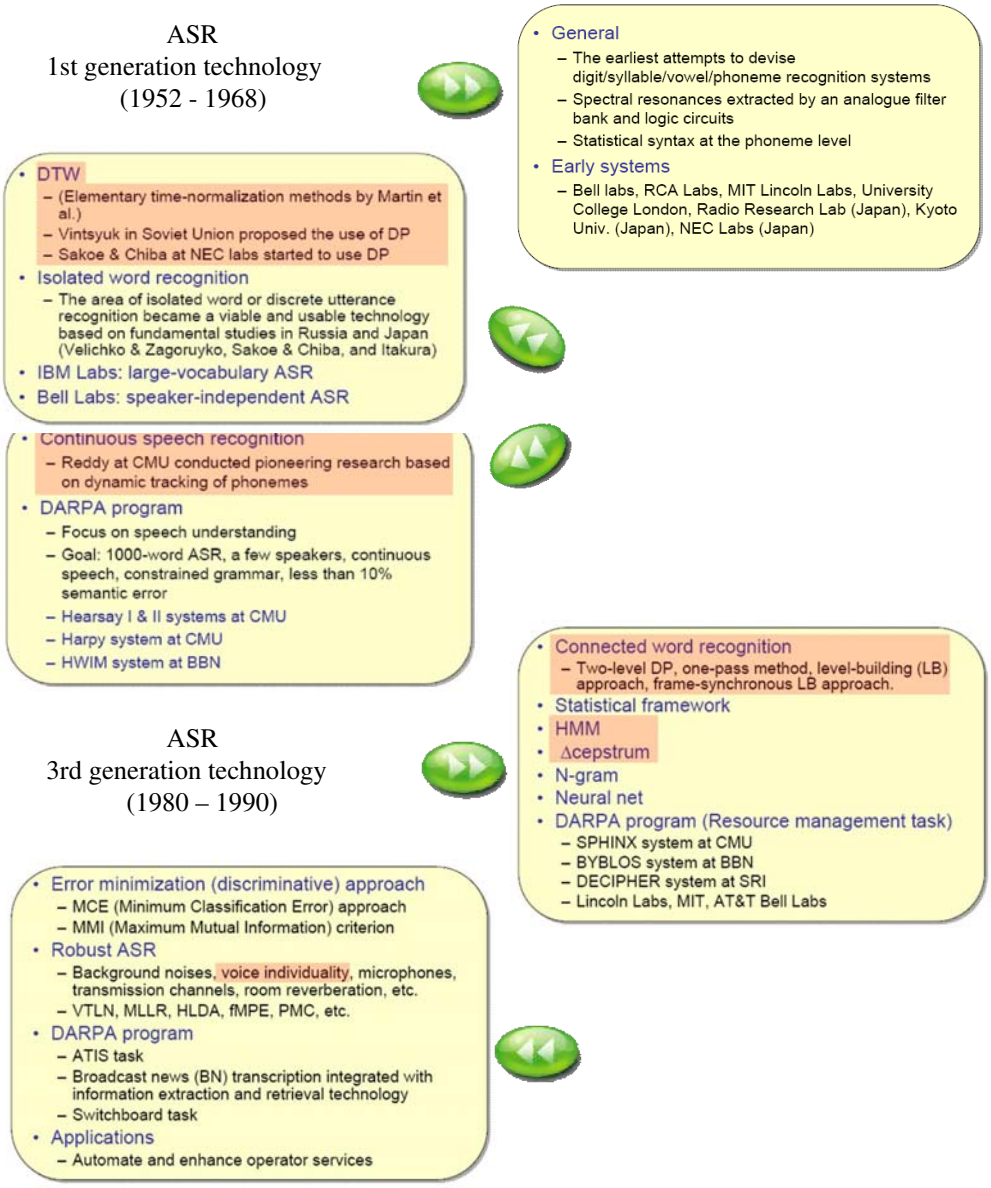


Figura 1: Apporti tecnologici più significativi influenzati dalla dimensione temporale del parlato nel campo del riconoscimento automatico

Nonostante enormi progressi, di seguito illustrati in Figura 2, manca ancora molto però all'utilizzazione diffusa di queste tecnologie soprattutto a causa della loro inaffidabilità quando vengono utilizzate realmente 'sul campo', quando cioè si devono superare i problemi relativi ad esempio al riconoscimento automatico del parlato in situazioni

rumorose (cocktail party, rumori sovrapposti, rumori di canale…) oppure al riconoscimento automatico di parlato spontaneo.

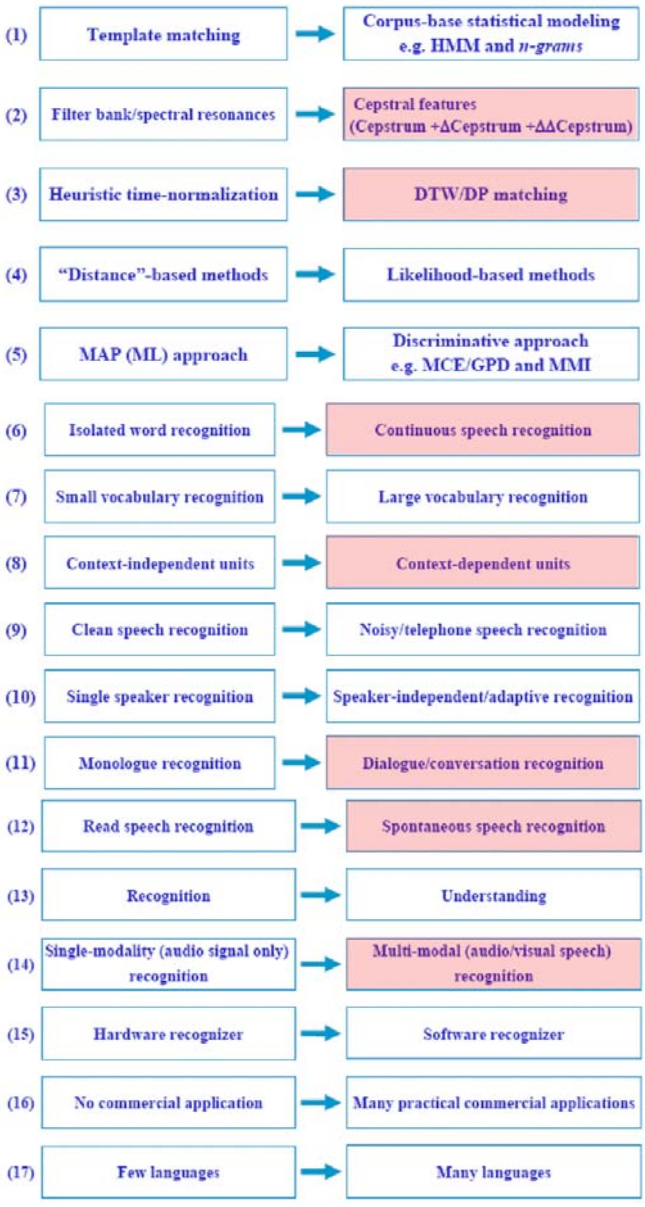


Figura 2: Principali innovazioni tecnologiche in cui la dimensione temporale gioca un ruolo fondamentale

A titolo di esempio si sottolineano gli enormi miglioramenti dovuti al passaggio fra l'utilizzazione di caratteristiche basaste sull'analisi effettuata da banchi di filtri in frequenza

e l'utilizzazione di caratteristiche basate sull'analisi del Cepstrum e delle relative velocità ed accelerazioni (Δ, ΔΔ), oppure fra l'utilizzazione della normalizzazione temporale euristica utilizzata agli inizi degli anni 70 per uniformare i confronti fra il parlato target e i modelli di parola memorizzati e l'utilizzazione della tecnica di programmazione dinamica (*Dynamic Time Warping*), oppure l'introduzione della modellizzazione basata sulla teoria delle catene di *Markov* nascoste.

Riassumendo, in Figura 3, è graficamente illustrato l'avvicendarsi temporale delle varie generazioni dei sistemi di riconoscimento., dalla preistoria (1920) agli anni recenti (3.5G). L'interrogativo fondamentale dei prossimi anni, per risolvere i problemi rimasti per un'utilizzazione diffusa di queste tecnologie è:

QUALI SARANNO LE CARATTERISTICHE DELLA QUARTA GENERAZIONE DEI SISTEMI ASR?

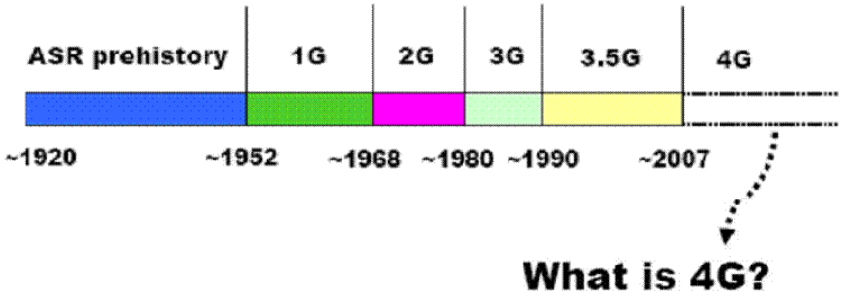


Figura 3: ASR dalla preistoria ai giorni nostri, dalla prima generazione alla terza, alla terza e mezzo ed alla futura quarta.

Le maggiori difficoltà che devono a tutt'oggi essere ancora risolte sono illustrate in Figura 4. In particolare le più rilevanti sono quelle relative alla variabile velocità di eloquio, all'estrema variabilità dell'accento, dello stile ed in generale della prosodia, tutte caratteristiche in cui la dimensione temporale risulta di fondamentale importanza.

In conclusione, per quanto riguarda il TAL, negli ultimi 50 anni sono stati fatti passi giganteschi e le maggiori innovazioni tecnologiche sono state focalizzate al miglioramento dei sistemi di riconoscimento soprattutto in termini di aumento della loro robustezza. Tuttavia il 60% (16/28) dei "problemi irrisolti" elencati da Beek *et al.* nel 1977 non sono ancora stati risolti.

Una comprensione assai più dettagliata del processo di produzione e percezione del parlato sarà necessariamente richiesta in futuro prima che i sistemi di riconoscimento vocale automatico possano avvicinarsi alla prestazione umana e di sicuro gran parte degli avanzamenti significativi in questo campo verranno dalla estesa collaborazione fra questa necessaria conoscenza e l'elaborazione della conoscenza basata invece sulle architetture e sulle teorie del riconoscimento di pattern basati sulla statistica.

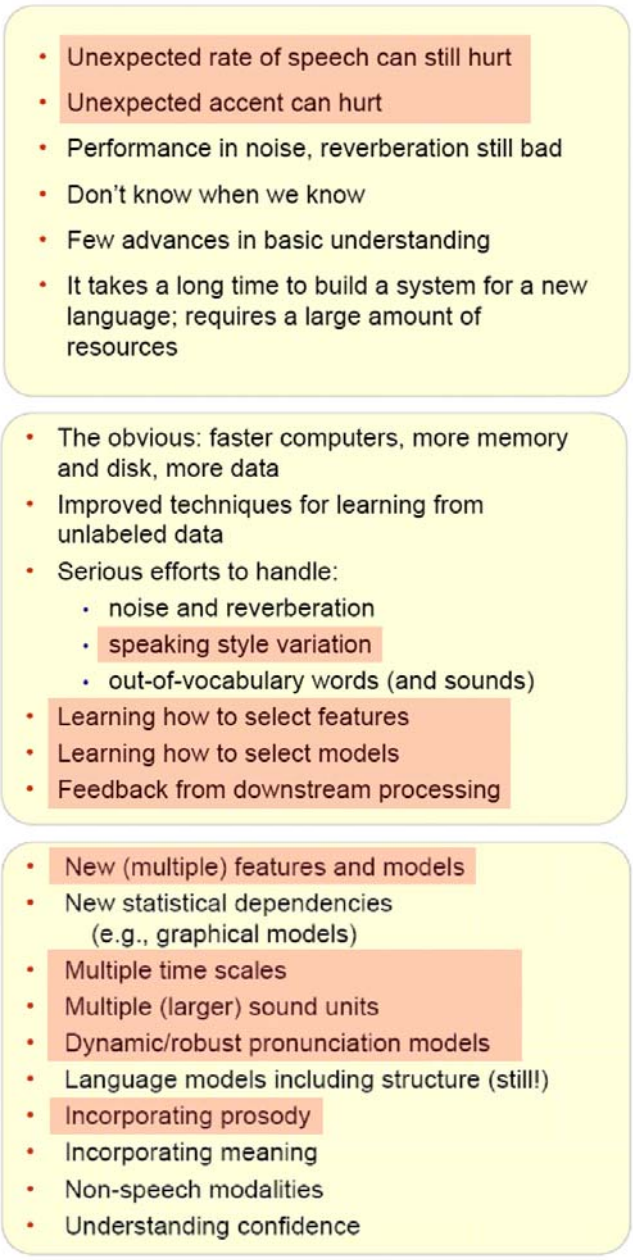


Figura 4: Elenco delle maggiori difficoltà che a tutt'oggi devono essere ancora risolte per una diffusione completa delle tecnologie TAL ed in particolare del riconoscimento automatico del parlato

4. DIMENSIONE TEMPORALE E ESPRESSIONI FACCIALI

Già nel 1921, Flach sosteneva che solo la dinamica di un movimento è non-ambigua e convincente: "only the dynamics of a movement is unambiguous and convincing". La configurazione delle azioni facciali (espressioni relative sia a specifiche emozioni sia ad unità di azione individuali) rispetto alle emozioni ed all'intenzione comunicativa è un importante tema di ricerca. Meno invece si conosce circa la sincronizzazione delle azioni facciali, anche perché la misurazione manuale della sincronizzazione è assai complicata e laboriosa. Tuttavia, sappiamo che (Cohn, 2007) noi siamo altamente sensibili alla sincronizzazione delle azioni facciali nelle interazioni sociali (Edwards, 1998). Le azioni facciali più lente (vedi Figura 5), ad esempio, sembrano essere più genuine e naturali (Krumhuber & Kappas, 2005), come pure lo sembrano essere quelle più sincrone nei loro movimenti (Frank & Ekman, 1997). In particolare, le espressioni facciali più sottili diventano visibili soltanto quando le informazioni di movimento sono a disposizione di chi le percepisce (Ambadar, Schooler & Cohn, 2005).

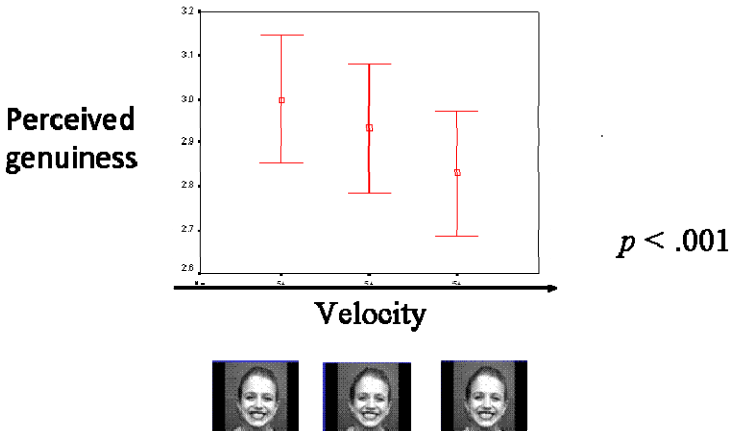


Figura 5: Risultati sperimentali a sostegno dell'ipotesi che la genuinità di un espressione (il sorriso in questo caso) è fortemente correlata alla lentezza della sua realizzazione articolatoria

La dinamica è cioè particolarmente importante per inferire l'intenzione comunicativa. Alcuni studi condotti dal gruppo di ricerca di CMU utilizzando tecniche automatiche di analisi di immagini facciali per misurare la sincronizzazione delle azioni facciali, hanno provato (vedi Figura 5) che le caratteristiche dinamiche riescono a discriminare fra i sorrisi intenzionali e quelli spontanei con un livello di precisione dell' 89% (Cohn & Schmidt, 2004). Usando caratteristiche simili, il divertimento, l'imbarazzo ed il sorriso 'gentile' sono stati discriminati con una precisione dell'83% (Kanade, Hu & Cohn, 2005), che è paragonabile a quella umana. Lavori più recenti suggeriscono inoltre che la coordinazione multimodale dell'espressione facciale, del movimento della testa e dei gesti sono caratteristiche specifiche dell'imbarazzo (Keltner, 1995).

5. OSSERVAZIONI CONCLUSIVE

L'unica osservazione che può essere fatta sulla base di queste brevi note è che senza una completa conoscenza della dimensione temporale dei meccanismi di produzione del parlato e, più in generale, nell'interpretazione di un qualsiasi atto comunicativo, una diffusione capillare e completa delle tecnologie TAL ed in particolare del riconoscimento automatico del parlato non potrà mai avvenire in maniera soddisfacente.

BIBLIOGRAFIA

Ambadar, Z., Schooler, J. & Cohn, J.F. (2005), Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial ex-pressions, *Psychological Science*, 16, 403-410.

Beek, B. (1977), An assessment of the technology of automatic speech recognition for military applications, *IEEE Trans. Acoustics, Speech, Signal Processing*, ASSP-25, 1977, 310-322.

Cohn, J.F. & Schmidt, K.L. (2004), The timing of facial motion in posed and sponta-neous smiles, *International Journal of Wavelets, Multiresolution and Information Processing*, 2, 1-12.

Cohn, J.F. (2007), Foundations of human-centered computing: Facial expression and emotion, *Proceedings of the International Joint Conference on Artificial Intelligence 2007,* Hyderabad, India, 5-12.

Edwards, K. (1998), The face of time: Temporal cues in facial expressions of emotion, *Psychological Science*, 9(4), 270-276.

Frank, M.G. & Ekman, P. (1997), The ability to detect deceit generalizes across different types of high-stakes lies, *Journal of Personality and Social Psychology*, 72, 1429-1439.

Furui, S. (2005), 50 years of progress in speech and speaker recognition, *Proceedings of 10th International Conference on Speech and Computer* 2005, Patras, Greece, October 17-19, 1-9.

Kanade, T., Hu, C. & Cohn, J.F. (2005), Facial expression analysis, Paper presented at the *IEEE International Workshop on Modeling and Analysis of Faces and Gestures*, Beijing, China, October 16, 2005.

Keltner, D. (1995), Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement and shame, *Journal of Personality and Social Psychology*, 68, 441-454.

Krumhuber, E. & Kappas, A. (2005), Moving smiles: The role of dynamic components for the perception of the genuineness of smiles, *Journal of Nonverbal Behavior*, 29, 3-24.

RECENTI SVILUPPI DI 'SONIC' PER L'ITALIANO: RICONOSCIMENTO AUTOMATICO DEL PARLATO INFANTILE

Piero Cosi
Istituto di Scienze e Tecnologie della Cognizione, Sede di Padova 'Fonetica e Dialettologia'
Consiglio Nazionale delle Ricerche
piero.cosi@pd.istc.cnr.it

1. SOMMARIO

In questo lavoro vengono descritti i risultati dei più recenti esperimenti di riconoscimento automatico di parlato infantile effettuati, mediante l'utilizzazione del sistema denominato SONIC, su un corpus di parlato letto da bambini di età compresa fra i 7 e i 13 anni. Il corpus utilizzato è stato raccolto presso alcune scuole del Trentino da parte dell'ITC-IRST ora FBK (Fondazione Bruno Kessler), nell'ambito di un progetto europeo denominato PF-STAR. In particolare, completando alcuni esperimenti realizzati passato, si è voluto integrare i nuovi modelli di riconoscimento allenati su voci di bambini nella versione italiana del *Colorado Literacy Tutor*. Il tasso di errore di riconoscimento iniziale di 15,1% per un insieme di 33 unità fonetiche (21,8% considerando un insieme di 40 unità fonetiche) è stato successivamente ridotto al 12,2% (18,6% considerando 40 unità) utilizzando una combinazione delle più aggiornate tecniche di adattamento comprendenti la normalizzazione di lunghezza del tratto vocale (*Vocal Tract Lenght Normalization*, VTLN), la normalizzazione della varianza dei coefficienti Cepstrali (*Cepstral coefficients Variance Normalization,* CVN) e l'utilizzazione di modelli fonetici addestrati in modalità indipendente dal parlante utilizzando le più recenti strategie iterative denominate *Structural MAP Linear Regression* (SMAPLR) e *Speaker Adaptive Training* (SAT). Questo lavoro è la continuazione ed il completamento naturale di un precedente simile lavoro (Cosi & Pellom, 2005) condotto su un insieme limitato dello stesso corpus di dati.

2. INTRODUZIONE

Il *Colorado Literacy Tutor* (CLT)[1] (Cole *et al.*, 2003), un sistema tecnologicamente avanzato ed interattivo costituito da una serie di *tool* computerizzati per l'insegnamento/apprendimento della lingua inglese e progettato sulla base delle più recenti teorie cognitive, mira a migliorare il livello di apprendimento degli studenti delle scuole primarie. Semplificando notevolmente CLT consiste di quattro moduli fortemente integrati fra loro denominati *Managed Learning Environment*, *Foundational Reading Skills Tutors*, *Interactive Books*, e *Latent Semantic Analysis (*LSA*)* e una caratteristica fondamentale è data dall'inserimento e dall'utilizzazione, nei moduli realizzati per l'apprendimento, delle più recenti e innovative tecnologie della comunicazione.

In particolare i 'libri interattivi' sono la piattaforma principale per la ricerca e lo sviluppo delle tecnologie sul linguaggio naturale e gli agenti animati. Incorporando, infatti, il riconoscimento automatico del parlato, il *Trattamento Automatico del Linguaggio* naturale (TAL) e le più recenti e innovative tecnologie grafiche di animazione al computer

[1] http://www.colit.org/

mirano a rendere sempre più naturale l'esperienza dell'apprendimento mediante ausili tecnologici.

In Cosi *et al.* (2004) vengono descritte le attività di ricerca iniziali rivolte allo sviluppo della versione italiana del CLT, l'*Italian Literacy Tutor* (*ILT*). L'ILT sarà realizzato basandosi su alcuni tool sviluppati in questi anni all'ISTC-CNR quali: la versione italiana di SONIC per il riconoscimento automatico del parlato infantile (Pellom, 2001; Pellom & Hacioglu 2003; Hagen *et al.*, 2003; Hagen *et al.*, 2004)[2], la versione italiana di FESTIVAL (Cosi *et al.*, 2001) per la sintesi da testo scritto e, LUCIA, una faccia parlante MPEG-4 (Cosi *et al.*, 2003) in grado di esprimersi emotivamente.

Parallelamente al CLT, l'ILT sarà costituito da una serie di *tool* computerizzati per l'insegnamento e l'apprendimento dell'italiano come lingua madre (L1) o lingua seconda (L2). Questo progetto, risultato della collaborazione fra Università, Centri di Ricerca e Scuole Pubbliche, mira a migliorare il livello e la qualità dell'apprendimento scolastico degli studenti delle scuole di primo livello, mediante l'utilizzo di un software educativo sviluppato per aiutare gli allievi ad imparare a leggere e a comprendere correttamente un testo scritto.

Questi *tool* di apprendimento hanno un'enorme potenzialità e possono essere utilizzati per:

- insegnare a leggere e a capire un testo, all'interno di un completo programma di lettura, cercando possibilmente di identificare in età precoce eventuali soggetti disabili;
- migliorare la qualità del processo di apprendimento degli allievi aiutandoli ad acquisire specifiche conoscenze ed abilità mediante una più efficace capacità di comprensione del testo e mediante nuove ed efficaci strategie di scrittura;
- insegnare una seconda lingua.

Una caratteristica fondamentale dell'*Italian Literacy Tutor* è quindi lo sviluppo di specifici strumenti per l'insegnamento della lettura agli allievi con particolari carenze. Molti bambini hanno infatti problemi di lettura che, qualora non vengano precocemente risolti, causano a lungo termine notevoli conseguenze negative. I soggetti che non riescono a leggere fluentemente generalmente sono impiegati in lavori secondari, sono caratterizzati da una notevole 'sottostima', e sono incapaci di raggiungere i risultati che la loro potenziale capacità intellettuale e creativa consentirebbe. Se i problemi di lettura sono diagnosticati in età prescolare o nei primi anni di scuola possono essere sicuramente superati e risolti. Sebbene sia ormai assodato che un programma specifico e personalizzato di sostegno alla lettura possa fornire un'efficace soluzione le scuole non hanno risorse sufficienti a soddisfare tutte le possibili richieste e necessità.

L'*Italian Literacy Tutor* è progettato per fornire una soluzione efficace a questo problema fornendo una serie di tool di apprendimento per migliorare le capacità di lettura degli allievi e per identificarne eventuali carenze. L'ILT integra due tipologie di strumenti per l'apprendimento, uno basato sulle tecnologie dell'animazione e del parlato, e l'altro basato sulle tecnologia della comprensione del linguaggio.

Il primo insieme di strumenti include i libri interattivi (*Interactive Books*) e i Tutors Lettori che sono concepiti per lavorare assieme all'interno di un programma comprensivo

[2] Il sistema di riconoscimento SONIC è liberamente disponibile per scopi di ricerca per merito dell'Università del Colorado (http://cslr.colorado.edu)

di lettura. I libri interattivi aiuteranno agli studenti ad imparare a riconoscere le parole, leggere velocemente e comprendere ciò che leggono. Essi forniscono un ambiente per apprendere che va da lettori principianti (che possono farsi raccontare le storie dai personaggi animati, e poi essere coinvolti in dialoghi con i personaggi per valutare e esercitare la comprensione), a lettori avanzati che sono in grado di leggere le storie e quindi di ricevere addestramento alla lettura. I libri interattivi serviranno ad individuare le abilità di lettura mancanti o deboli, e indicheranno i Tutors Lettori individualizzati che valuteranno e insegneranno queste abilità. Sono in fase di sviluppo una serie di esercizi progettati per insegnare le abilità di base (*Foundational Skills*) fondamentali per una corretta lettura del testo. Mediante questi esercizi gli allievi interagiscono con agenti animati per l'apprendimento delle nozioni di base di una determinata lingua (L1, L2) quali ad esempio la conoscenza dell'alfabeto, la discriminazione e la produzione dei suoni linguistici, la consapevolezza fonologica, il 'suono' e le parole, la struttura sillabica (Figura 2). L'apprendimento di queste capacità si è dimostrato particolarmente efficace nell'insegnamento della lettura in soggetti dotati di particolari problemi.

3. SISTEMA DI RICONOSCIMENTO AUTOMATICO PER IL PARLATO INFANTILE

Come già detto precedentemente ILT utilizza il sistema di riconoscimento del parlato continuo denominato SONIC, sviluppato all'università del Colorado, come architettura di base per il riconoscimento in tempo reale del parlato infantile (Pellom, 2001; Pellom & Hacioglu, 2003; Hagen *et al.*, 2003; Hagen *et al.*, 2004). Il riconoscitore implementa un'efficace strategia di ricerca (*time-synchronous, beam-pruned Viterbi token-passing*) mediante un albero rientrante statico di prefissi lessicali e utilizza modelli markoviani nascosti con misture di gaussiane a densità di probabilità continua, anche fra le parole (HMMs). A livello di *front-end* acustico il riconoscitore utilizza come vettore di informazione i coefficienti cepstrali PMDVR (Yapanel & Hansen, 2003) o quelli classici MFCC.

3.1. Dati di addestramento e porting *iniziale per l'italiano*

La versione per l'inglese americano di SONIC, utilizzata nel CLT, è stata allenata utilizzando il parlato di più di 1800 bambini fra gli 8 e I 15 anni per un totale di più di 50 ore di parlato per l'addestramento del sistema (Hagen *et al.*, 2003; Shobaki *et al.*, 2000). In un task di lettura ad alta voce, si è ottenuto un errore di riconoscimento rispettivamente dell'8% e dell'11,5% a seconda dell'implementazione off-line e real-time del sistema (Hagen *et al.*, 2004).

Per l'apprendimento della versione italiana del riconoscitore di parlato infantile è stata utilizzata la versione completa del corpus *ChildIt* realizzato dall'ITC-IRST (ora Fondazione Bruno Kessler – FBK) (Gerosa *et al.*, 2007) che è stato realizzato con le registrazioni di 171 bambini (85 femmine e 86 maschi) di età compresa fra i 7 e i 13 anni, nativi del Trentino.

Per ogni bambino sono state registrate approssimativamente circa 50-60 semplici frasi lette da alcuni libri adeguati alla loro età. Seguendo il lavoro di Gerosa *et al.* (2007), il corpus è stato diviso in un insieme di training di 129 bambini (64 femmine e 65 maschi) e un insieme di test di 42 bambini (21 femmine e 21 maschi) bilanciati per sesso ed età fra i 7 e i 13 anni. Le frasi di training e di test contenenti parole mal pronunciate o forti rumori sovrapposti sono state preventivamente escluse negli esperimenti che verranno descritti di seguito, mentre tutte le altre frasi, anche quelle annotate con fenomeni extra linguistici tipo

rumori dovuti ai parlanti (respiri, risate o colpi di tosse, ecc.), rumori generici non sovrapposti con il segnale vocale (rumore generico, parlato estraneo non trascritto) e suoni non verbali o pause piene, sono state incluse e solo la loro trascrizione fonetica derivata dalla corrispondente trascrizione ortografica è stata utilizzata in fase di training e test.

Il sistema di riconoscimento SONIC dell'Università del Colorado allenato per voci di adulti americani (16 kHz, parlato microfonico) è stato trasformato nella versione per voci infantili italiane nel modo seguente: in un prima fase si è determinato una mappatura fonetica fra i fonemi target italiani, considerando 40 unità, e quelli inglesi americani; successivamente, questa mappatura fonetica, è stata utilizzata per fornire un primo allineamento forzato ottenuto mediante algoritmo di *Viterbi* con i moduli di riconoscimento per l'inglese americano e questo allineamento è servito come *boot-strap* per il training vero e proprio dei modelli acustici per l'italiano. I fonemi target italiani e la loro mappatura con quelli inglesi americani è illustrata in Tabella 1.

IT	US	Example	IT	US	Example
i	IY	p**i**ni	i1	IY	così
E	EH	asp**e**tto	E1	EH	caff**è**
o	OW	p**o**lso	o1	OW	R**o**ma
u	UW	p**u**nta	u1	UW	più
k	K	**c**aldo	g	G	**g**atto
t	T	**t**orre	d	D	**d**ente
tS	TS	pe**c**e	dZ	JH	ma**g**ia
ng	NG	a**n**gora	nf	NG	a**n**fora
l	L	pa**l**o	r	R	**r**emo
s	S	**s**ole	z	Z	pe**s**o
e	EY	v**e**lo	e1	EY	merc**é**
a	AA	v**a**i	a1	AA	bont**à**
O	AW	c**o**sa	O1	AW	per**ò**
j	Y	p**i**ume	w	W	q**u**ale
p	P	**p**era	b	B	**b**otte
ts	TS	pi**zz**a	dz	ZH	**z**ero
m	M	**m**ano	n	N	**n**ave
J	N	le**gn**a	L	L	Sog**l**ia
f	F	**f**aro	v	V	**v**ia
S	SH	**Sc**i	SIL	SIL	silence

Tabella 1: insieme di fonemi (SAMPA) utilizzati per il riconoscimento di parlato infantile italiano e corrispondente mappatura sull'inglese americano per il bootstrapping del sistema

Mediante la mappatura fonetica, la trascrizione ortografica delle frasi target e un lessico di pronuncia, il sistema effettua inizialmente l'allineamento forzato mediante algoritmo di *Viterbi* delle frasi di training fornendo al riconoscitore l'associazione fra i *frames* acustici e gli stati dei modelli di *Markov* nascosti (HMM) associati alle parole delle frasi di training. Negli esperimenti che seguono, ogni fonema è rappresentato da un modello HMM a tre stati. Una volta determinato l'allineamento, vengono stimati i modelli HMM sulla base di alberi di decisione binari (*decision-tree state-clustered triphone* HMM). In SONIC, le domande che vengono poste nell'albero di decisione binario possono essere formulate in

modo automatico per massimizzare la verosimiglianza (*likelihood*) dell'insieme di dati di training e non sono quindi necessarie domande basate su un'approfondita conoscenza linguistica per il *porting* di una lingua su di un'altra.

Ad ogni stato sono state assegnate da 6 a 24 misture di Gaussiane a seconda del materiale di training disponibile e, una volta allenati i modelli acustici iniziali, si è proceduto poi sequenzialmente ad un successivo allineamento forzato di *Viterbi* ed ad un nuovo addestramento per migliorare via-via i modelli acustici finali. Nei paragrafi seguenti sono descritti una serie di esperimenti che ben illustrano le problematiche incontrate nello sviluppo di un sistema di riconoscimento di parlato infantile

3.2 Corpus ChildIt

Gli esperimenti di riconoscimento fonetico sono stati eseguiti sul corpus *ChildIt* facendo uso di 42 parlanti per il test del sistema ovviamente non inclusi in quelli utilizzati per l'addestramento dei moduli acustici. Per il riconoscimento fonetico l'insieme di fonemi utilizzato è consistito di 40 unità acustiche primarie (AUs) (vedi la Tabella 1). I risultati per il riconoscimento fonetico sono presentati facendo uso di questo insieme di 40 unità come pure di un insieme ridotto di 33 unità acustiche che non considera gli eventuali errori di riconoscimento riscontrati sulle vocali accentate o atone (per esempio, 'a' con 'a1', oppure 'o' con 'o1'), errori che non pregiudicherebbero la *performance* del sistema qualora si utilizzassero le parole, quali unità di riconoscimento, invece delle unità acustiche assieme ad uno specifico modello del linguaggio.

3.3 Esperimenti

In ogni esperimento sono state utilizzate le sequenze fonetiche ottenute tramite l'allineamento di *Viterbi* della trascrizione ortografica dei dati di test come trascrizione fonetica di riferimento. Il modulo per l'allineamento fonetico forzato realizzato all'interno di SONIC tiene in considerazione, oltre a selezionare la migliore pronuncia per una parola dato un insieme di pronunce alternative estratte da un dizionario lessicale italiano, anche del rilevamento e inserzione automatici del simbolo della pausa o silenzio. Ovviamente, nella migliore delle ipotesi, sarebbe preferibile disporre di un corpus etichettato e segmentato manualmente a livello fonetico per considerare eventuali inserzioni, cancellazioni e sostituzioni di unità fonetiche nella realizzazione effettiva delle frasi target sia per il training che per il test. Per ognuno degli esperimenti descritti nei paragrafi seguenti, è stato inoltre stimato un modello fonetico del linguaggio a tri-grammi (Clarkson & Rosenfeld, 1997) a partire dalle sequenze fonetiche risultanti dai dati allineati di addestramento che consistono di 13765 espressioni.

3.3.1. Riconoscimento di parlato infantile con modelli acustici di parlato adulto

Nella nostra prima serie di esperimenti, si desiderava capire il tasso di errore fonetico di riconoscimento di un sistema mal adattato, cioè, un sistema addestrato su voci di parlanti adulti per riconoscere il parlato infantile. Si desiderava inoltre quantificare la riduzione degli errori che poteva essere ottenuta qualora si fossero utilizzati alcuni metodi di normalizzazione e di adattamento. Per questo esperimento sono stati utilizzati i modelli acustici per l'italiano addestrati su parlanti adulti mediante il corpus APASCI realizzato da FBK (ex ITC-irst). APASCI è un corpus di parlato italiano adulto registrato in una camera silente mediante microfono Sennheiser MKH 416 T. Il corpus contiene 5.290 frasi foneticamente ricche oltre a 10.800 cifre isolate (più di 10 ore di parlato). Il materiale vocale è stato letto da 100 parlanti (50 maschi e 50 femmine) italiani. E' stata utilizzata la procedura di

porting dall'inglese americano all'italiano descritta nel paragrafo 3.1 e sono stati stimati i modelli acustici indipendenti dal parlante e quelli dipendenti dal genere maschile o femminile. Per ridurre il disallineamento fra i modelli di parlato adulto ed i dati acustici dei bambini, è stata applicata la regressione lineare strutturata massima a posteriori (*Maximum-A-Posteriori, MAP*) non supervisionata (*SMAPLR*), utilizzando l'uscita fonetica del riconoscitore opportunamente pesata mediante una misura di affidabilità (Siohan, 2002). Le medie e le varianze delle gaussiane del sistema sono state adattate utilizzando la procedura *SMAPLR* dopo ogni passaggio di decodifica e sono state utilizzate per ottenere un risultato fonetico migliore dil riconoscimento. I risultati sono indicati in Tabella 2 (a).

Ricerche precedenti inoltre hanno indicato che la normalizzazione della lunghezza del tratto vocale (*Vocal Tract Lenght Normalization,* VTLN) mediante deformazione dell'asse delle frequenze prima dell'estrazione del vettore delle caratteristiche acustiche può sicuramente essere di aiuto per la riduzione del non allineamento fra il parlato dei bambini ed i modelli acustici per gli adulti. In SONIC è implementato il metodo di deformazione dell'asse delle frequenze descritto in Welling *et al.* (1999). La funzione VTLN determina il fattore di deformazione, necessario per far corrispondere i dati anatomici medi di adulti comparati a quelli dei bambini, che varia fra 0,88 e 1,12 per ogni parlante in modo tale da massimizzare la verosimiglianza (likelihood) dei dati di test. I risultati degli esperimenti che combinano SMAPLR e VTLN sono riassunti in Tabella 2(b). Dalla tabella 2(a) possiamo vedere che il tasso di errore fonetico iniziale è 39,2% per un sistema che consiste di 40 unità acustiche (AU) (31,1% per 33 AU) quando i modelli acustici addestrati su voci di adulti sono stati utilizzati per riconoscere il parlato infantile. Come ci si poteva attendere, inoltre, i modelli acustici addestrati unicamente su parlanti adulti femmine, forniscono un piccolo miglioramento rispetto a quelli indipendenti dal parlante – riducendo il tasso di errore fonetico iniziale a 36,8% e a 28,7% rispettivamente per 40 o 33 AU. L'adattamento mediante SMAPLR riduce ulteriormente il tasso di errore fonetico a 28,1% e 20,7% (vedi la Tabella 2(a) rispettivamente per 40 o 33 unità acustiche.

La combinazione di VTLN nello spazio delle caratteristiche acustiche con SMAPLR nello spazio del modello riduce ancora il tasso di errore a 26,7% e a 19,3% (Tabella 2(b)). Riassumendo, può essere raggiunta una riduzione dell'errore relativo di quasi 32% combinando l'adattamento dello spazio-acustico (SMAPLR) e l'adattamento dello spazio delle caratteristiche (VTLN) ai modelli acustici addestrati esclusivamente su voci femminili adulte.

SMAPLR Adaptation	**(a) Speaker Ind.**		**(b) Adult Female**	
	PER 40 AU	**PER 33 AU**	**PER 40 AU**	**PER 33 AU**
First-Pass	39,2%	31,1%	36,8%	28,7%
+Adapt Iter. 1	31,7%	24,1%	29,6%	22,0%
+Adapt Iter. 2	29,7%	22,2%	27,8%	20,3%
+Adapt Iter. 3	28,9%	21,5%	27,0%	19,6%
+Adapt Iter. 4	28,4%	21,0%	26,5%	19,1%
+Adapt Iter. 5	28,1%	20,7%	26,5%	18,8%

Tabella 2: Tasso di errore fonetico per il riconoscimento del parlato di bambini (*PER*) in funzione della ripetizione dell'addestramento SMAPLR per i modelli acustici addestrati su voci di parlanti adulti indipendenti dal parlante (a) e di parlanti femminili (b) con adattamento VTLN e SMAPLR

Nella sezione seguente, viene descritto lo sviluppo dei modelli acustici addestrati unicamente sul parlato di bambini per illustrare il grado di non allineamento che tuttora esiste fra i modelli adulti adattati alle voci di bambini ed i modelli infantili veri e propri.

3.3.2 Addestramento di Viterbi dei modelli di parlato infantile italiano

Come citato nella sezione 3.1, il metodo di *porting* di SONIC dall'inglese all'italiano fa affidamento su di un iniziale mappatura fonetica basata su una conoscenza linguistica specifica della differenza fra i fonemi target e i fonemi della lingua di partenza. Fino ad oggi, SONIC è stato utilizzato per il *porting* su quasi 20 lingue e l'esperienza ha indicato che l¨'accuratezza della mappatura iniziale ha un impatto minimo sul tasso di errore finale dei modelli acustici risultanti. In questo lavoro sono stati effettuati un numero totale di 6 allineamenti di *Viterbi* e di re-training dei modelli acustici per ottenere i modelli finali di riconoscimento di voci di bambini. In Tabella 3 è illustrato il tasso di errore di riconoscimento fonetico in funzione della ripetizione dell'allineamento ed è chiaro che 6 passaggi di allineamento sono sufficienti per raggiungere la convergenza del sistema.

Viterbi Training Step	**Children's Acoustic Models**	
	PER (40 AU)	**PER (33 AU)**
Align / Train Pass 0	24,4%	17,4%
Align / Train Pass 1	22,8%	15,9%
Align / Train Pass 2	22,1%	15,4%
Align / Train Pass 3	21,7%	15,1%
Align / Train Pass 4	21,7%	15,1%
Align / Train Pass 5	21,7%	15,0%
Align / Train Pass 6	21,8%	15,1%

Tabella 3: Tasso di errore sul riconoscimento fonetico (PER) in funzione dell'iterazione dell'allineamento di Viterbi per l'addestramento del modulo acustico di riconoscimento di parlato infantile sul corpus di parlato letto *ChildIt* dell'ITC-irst (ora FBK)[3]

Vale la pena di notare che i modelli di riconoscimento base per i bambini rivelano una riduzione del 10% del tasso di errore del riconoscimento fonetico se confrontato ai migliori modelli adulti adattati al parlato dei bambini (vedi Tabella 2).

3.3.3 Esperimenti di riconoscimento con modelli acustici di parlato infantile

Come descritto nella sezione 3.3.1, in modo simile a quello eseguito per gli esperimenti con i modelli allenati su parlato adulto, i modelli base allenati direttamente sulle voci di bambini sono stati estesi mediante l'introduzione dell'adattamento iterativo SMAPLR. I risultati di questo esperimento sono indicati in Tabella 4(a).

[3] Si noti che all'iterazione iniziale (0) l'allineamento è ottenuto utilizzando i modelli acustici inglesi americani mentre nelle iterazioni 1-6 è ottenuto con i modelli acustici addestrati direttamente sul parlato di bambini.

Children's Speech Phonetic Recognition	(a) SMAPLR Adaptation		(b) SMAPLR & VTLN	
	PER 40 AU	**PER 33 AU**	**PER 40 AU**	**PER 33 AU**
First-Pass	21,8%	15,1%	21,8%	15,1%
+Adapt Iter. 1	20,3%	13,6%	19,0%	12,6%
+Adapt Iter. 2	19,9%	13,3%	18,7%	12,4%
+Adapt Iter. 3	19,8%	13,2%	18,7%	12,3%
+Adapt Iter. 4	19,8%	12,3%	18,7%	12,3%
+Adapt Iter. 5	19,8%	13,2%	18,7%	12,3%

Tabella 4: (a) tasso di errore di riconoscimento fonetico (PER) in funzione della ripetizione dell'adattamento SMAPLR; (b) tasso di errore di riconoscimento fonetico (PER) in funzione della ripetizione dell'adattamento SMAPLR a cui è stata aggiunta la normalizzazione del tratto vocale (VTLN)

Diversamente dagli esperimenti descritti in 3.3.1 con i modelli adulti, sono necessarie meno ripetizioni di adattamento per raggiungere il tasso di errore di riconoscimento fonetico più basso. L'adattamento acustico applicato ai modelli dei bambini riduce ulteriormente l'errore di quasi il 9%. Inoltre, come visto in precedenza, si voleva verificare se sarebbe stato possibile ottenere un miglioramento del sistema qualora le differenze del tratto vocale fossero rimosse fra i vari bambini nell'insieme di training e per far questo è stata applicata la normalizzazione del tratto vocale (VTLN) per ogni bambino appartenente all'insieme di training ed è stata applicata anche la normalizzazione VTLN previa stima del fattore di ditorsione dell'asse delle frequenze per ogni bambino appartenente all'insieme di test (Welling *et al.,* 1999). Possiamo vedere dalla Tabella 4 che incorporando la procedura VTLN il tasso di errore di riconoscimento fonetico si riduce dal 21,8% al 18,7% per il sistema con 40 unita acustiche e dal 15,1% al 12,3% per il sistema ridotto con 33 unità acustiche. Come già anticipato, i guadagni dovuti all'applicazione della procedura di *VTLN* sono meno sostanziali in questa condizione di quanto non lo siano in condizioni di non allineamento sostanziale dei modelli (cioè, modello adulto con parlato infantile).

Da ultimo, la tecnica di adattamento denominata *Speaker Adaptive Training* (SAT) mira a rimuovere le caratteristiche specifiche di un parlante sui dati di training al fine di stimare i parametri dei modelli acustici indipendenti dai parlanti. In SONIC la procedura SAT viene realizzata mediante la stima di una singola trasformazione lineare nello spazio delle caratteristiche acustiche per ogni parlante dell'insieme di addestramento.

Questa funzione di trasformazione viene stimata con il vincolo di massimizzare la probabilità dei dati di addestramento una volta stimato il modello acustico dopo aver applicato la procedura di VTLN.

Durante la fase di test, il fattore di distorsione viene stimato mediante un unica funzione di trasformazione (*Maximum Likelihood Linear Regression*) nello spazio delle caratteristiche acustiche prima del riconoscimento. Questo sistema finale riduce il PER dal 21,8% al 18,6% per il sistema con 40 unità acustiche e dal 15,1% al 12,2% per il sistema ridotto con 33 unità, come illustrato in Tabella 5.

Italian Children's Speech Phonetic Recognition	SMAPLR + VTLN + SAT	
	PER 40 AU	PER 33 AU
First-Pass	21,8%	15,1%
+Adapt Iter. 1	19,0%	12,5%
+Adapt Iter. 2	18,7%	12,3%
+Adapt Iter. 3	18,6%	12,2%
+Adapt Iter. 4	18,6%	12,2%
+Adapt Iter. 5	18,6%	12,2%

Tabella 5: Errore di riconoscimento fonetico (PER) con modelli acustici di parlato infantile per un sistema in cui sono state applicate le procedure SMAPLR, VTLN e *Speaker Adaptive Training* (SAT) in funzione dell'iterazione di *re-training*

3.3.4. Discussione degli esperimenti

Mentre il tasso di errore del sistema allenato su voci di bambini è paragonabile e addirittura migliore di quello ottenuto da sistemi simili sullo stesso corpus (ad esempio paragonabile al 22,7% ottenuto da un sistema analogo con 28 unità fonetiche come quello utilizzato in Giuliani & Gerosa, 2003), esiste ancora un significativo margine di miglioramento per un sistema che utilizzi modelli acustici allenati su parlato adulto e utilizzati per decodificare parlato infantile. Infatti quando sono state applicate entrambe le tecniche VTLN e SMAPLR in una condizione di disallineamento adulti/bambini il sistema finale ha ottenuto un tasso di errore fonetico del 19,3% dimostrando di ridurre l'errore fonetico iniziale del 28%. Ciò nonostante, persiste ancora un notevole 30% di differenza relativa fra l'utilizzazione di modelli acustici allenati su parlato adulto e modelli acustici allenati su parlato infantile per la decodifica di quest'ultimo.

Il tasso di errore di riconoscimento iniziale di 15,1% per un insieme di 33 unità fonetiche (21,8% considerando un insieme di 40 unità fonetiche) è stato successivamente ridotto al 12,2% (18,6% considerando 40 unità) utilizzando una combinazione delle più aggiornate tecniche di adattamento comprendenti la normalizzazione di lunghezza del tratto vocale (VTLN), la normalizzazione della varianza dei coefficienti Cepstrali (*Cepstral coefficients Variance Normalization,* CVN) e l'utilizzazione di modelli fonetici addestrati in modalità indipendente dal parlante utilizzando le più recenti strategie iterative denominate *Structural MAP Linear Regression* (SMAPLR) e *Speaker Adaptive Training* (SAT).

4. CONCLUSIONI E SVILUPPI FUTURI

Lo sviluppo di sistemi di riconoscimento di parlato infantile spesso si presenta come un compito di ardua soluzione a causa della spesso totale mancanza di risorse acustiche utilizzabili per l'allenamento dei modelli acustici. In questo lavoro, il sistema di riconoscimento denominato *SONIC* e sviluppato per l'inglese è stato adattato all'italiano ed in particolare è stato considerato il caso del parlato infantile di bambini compresi nella fascia di età compresa fra i 7 e i 13 anni.

Questi nuovi modelli acustici per il parlato infantile italiano sono stati incorporati nel *Colorado Literacy Tutor* (CLT), sviluppato al *Centre for Speech and Language Research* (CSLR) della *University of Colorado di Boulder*, per la lingua inglese, quale primo passo

verso lo sviluppo della sua corrispondente versione italiana l'*Italian Literacy Tutor* (Cosi *et al.*, 2004), un sistema interattivo e personalizzato per l'apprendimento della lingua italiana. In particolare in Figura 1 è illustrata la videata iniziale di un libro interattivo utilizzato per insegnamento/apprendimento della lettura in italiano, in cui si può notare in alto a destra l'assistente virtuale LUCIA (Cosi *et al.*, 2003) che può ad esempio leggere il testo ed in generale interagire con gli utenti durante il processo di apprendimento.

All'interno del CLT e in futuro dell'ILT, il riconoscimento del parlato viene utilizzato per seguire il livello raggiunto dal bambino nell'apprendimento della lettura, e può assisterlo nella rilevazione degli eventuali possibili errori, in generale allo scopo di fornire alcune informazioni utilizzabili in fase di misurazione della fluenza di lettura. Per migliorare il riconoscimento dei testi il sistema di riconoscimento costruisce un modello del linguaggio del testo del libro a tri-grammi al fine di fornire quella flessibilità necessaria per inserire/cancellare/sostituire le parole in funzione del modello acustico elaborato e per fornire inoltre delle misure di confidenza acustica calcolate dal lattice di ipotesi di parola fornito dal riconoscitore.

Quando il bambino parla, le ipotesi parziali vengono inviate al modulo di riconoscimento che determina la posizione attuale della lettura allineando ogni ipotesi parziale con la storia del testo utilizzando un algoritmo di ricerca basato sulla programmazione dinamica. Hagen *et al.* (2004) descrive in dettaglio gli avanzamenti più recenti raggiunti sia nella modellizzazione acustica sia nella realizzazione di efficienti modelli del linguaggio per il riconoscimento della lettura di parlato infantile.

Di conseguenza, ILT utilizzerà in futuro un modulo per il controllo della storia/ cronologia delle parole anche a cavallo di frase, nuovi modelli del linguaggio a trigrammi dinamici e adattabili alla posizione nel testo, come pure, a livello acustico, un modulo specifico per la normalizzazione del tratto vocale, per l'addestramento adattativo ai parlanti, e per l'addestramento dei parlanti senza supervisione.

Il sistema risultante per l'inglese americano ha dimostrato di raggiungere un tasso di errore globale di riconoscimento di parola dell'8,0%. In uno studio successivo, gli errori fatti da questo sistema di base sono stati analizzati ed utilizzati per migliorare lo sviluppo di un sistema in grado di riconoscere possibili errori di lettura (Lee *et al.*, 2004).

Basandosi su tutte queste ricerche, il sistema di riconoscimento SONIC è stato esteso per realizzare il tracciamento della lettura e l'analisi del segnale verbale utilizzando come unità acustiche delle unità di dimensioni più piccole della parola (Hagen & Pellom, 2005).

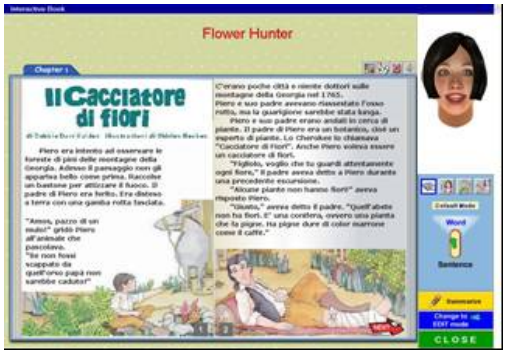


Figura 1: *Italian Literacy Tutor* (ILT), videata iniziale di un libro interattivo

RINGRAZIAMENTI

Un calorosissimo ringraziamento va all'intero gruppo di ricerca CSLR (*Computer Spoken Language Research*) dell'Università del Colorado e soprattutto a Bryan Pellom (ora in *Rosetta Stone*) per i suoi insostituibili suggerimenti e la sincera amicizia.

5. BIBLIOGRAFIA

Cole R., van Vuuren, S., Pellom B. *et al.* (2003), Perceptive Animated Interfaces: First Steps Toward a New Paradigm for Human Computer Interaction, in *Proceedings of the IEEE*, 91, September 2003, 1391-1405.

Cosi, P., Delmonte, R., Biscetti, S., Cole, R., Pellom, B. & van Vuuren, S. (2004), Italian Literacy Tutor: tools and technologies for individuals with cognitive disabilities, in *Proceedings InSTIL/ICALL Symposium*, Venice, Italy.

Cosi, P. & Pellom, B. (2005), Italian Children's Speech Recognition For Advanced Interactive Literacy Tutors, in *CD-Rom Proceedings INTERSPEECH 2005*, Lisbon, Portugal, 2201-2204.

Gerosa, M., Giuliani, D. & Brugnara, F. (2007), Acoustic Variability and automatic recognition of children's speech, *Speech Communication*, 49, 847-860.

Giuliani, D. & Gerosa, M. (2003), Investigating Recognition of Children's Speech, in *Proceedings of the ICASSP*, Hong Kong, 2003.

Hagen, A. & Pellom, B. (2005), A Multi-Layered Lexical-Tree Based Token Passing Architecture for Efficient Recognition of Subword Speech Units, in *2nd Language & Technology Conference*, April 2005, Poznan, Poland.

Hagen, A., Pellom, B. & Cole ,R. (2003), Children's Speech Recognition with Application to Interactive Books and Tutors, in *Proceedings of the ASRU*, St. Thomas, USA, 2003.

Hagen, A., Pellom, B., Van Vuuren, S. & Cole, R. (2004), Advances in Children's Speech Recognition within an Interactive Literacy Tutor, in *Proceedings of the HLT-NAACL*, Boston Massachusetts, USA, 2004.

Lee, K., Hagen, A., Romanyshyn, N., Martin, S. & Pellom, B. (2004), Analysis and Detection of Reading Miscues for Interactive Literacy Tutors, in *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, 2004.

Pellom, B. (2001), *SONIC: The University of Colorado Continuous Speech Recognizer*, Technical Report TR-CSLR-2001-01, University of Colorado, USA, 2001.

Pellom, B. & Hacioglu, K. (2003), Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task, in *Proceedings of the ICASSP*, Hong Kong, 2003.

Siohan, O., Myrvoll, T. & Lee, C.H. (2002), Structural Maximum a Posteriori Linear Regression for Fast HMM Adaptation, *Computer, Speech and Language*, 16, 5-24.

Welling, L., Kanthak ,S., Ney H. (1999), Improved Methods for Vocal Tract Length Normalization, in *Proceedings of the ICASSP*, Phoenix Arizona, 1999.

Yapanel, U.H., Hansen, J.H.L. (2003), A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition, in *Proceedings EUROSPEECH 2003*, Geneva, Switzerland, September 1-4, 1281-1284.

TEST FONETICO DELLA PRIMA INFANZIA PER BAMBINI DAI 18 AI 36 MESI: ANALISI CON 'PHON' DEI PRIMI DATI RACCOLTI

Claudio Zmarich [a], Serena Bonifacio [b], Maria Pia Bardozzetti [a], Caterina Pisciotta [a],
[a] Istituto di Scienze e Tecnologie della Cognizione (ISTC), C.N.R., Sede di Padova;
[b] IRCCS 'Burlo Garofolo', Trieste
claudio.zmarich@pd.istc.cnr.it, logopedia@burlo.trieste.it, mariapia.bardozzetti@libero.it, caterina.pisciotta@pd.istc.cnr.it

1. SOMMARIO

Questo articolo si propone il duplice obiettivo di descrivere un nuovo test per la valutazione delle capacità fonetico-fonologiche di bambini dai 18 ai 36 mesi, e un nuovo programma, *Phon*, per la codifica e l'analisi semiautomatica degli aspetti segmentali del parlato. Il nesso tra i due oggetti dell'articolo sarà dato dall'applicazione di *Phon* ai primi dati raccolti tramite il *Test Fonetico per la Prima Infanzia*.

Il test, progettato da Serena Bonifacio e Claudio Zmarich, è ancora in fase di sviluppo, e la versione che qui viene descritta risale alla fine del 2007. Attualmente in Italia non si dispone di uno strumento clinico che permetta di valutare le capacità fonetiche in età così precoci, sebbene il test PFLI (Bortolini, 1995) si proponga di farlo dai 24 mesi in poi. Questo nuovo test si propone la stesura dell'inventario fonetico, basato sulla produzione verbale del bambino (stimolata ma non ripetuta), per accertare se un fono o un gruppo consonantico risultano acquisiti in sede iniziale e non iniziale di parola (cfr. Zmarich & Bonifacio, 2005). Il test è composto da una lista di parole suddivise per fascia d'età, che rappresentano gli item lessicali più prodotti dai bambini per quella fascia (Caselli & Casadio, 1995; Caselli, Pasqualetti & Stefanini, 2007). Le parole sono state scelte in base a un criterio fonetico, e cioè che tutti i foni della lingua italiana siano attestati in almeno 3 parole per ciascuna posizione lessicale (vedi anche Zmarich *et al.*, accettato per la pubblicazione).

Sono state analizzate, in questo studio pilota, le produzioni verbali audioregistrate di 12 bambini con sviluppo linguistico tipico e cioè con dimensione del vocabolario espressivo > al 5° percentile, rilevato con il PVB (Caselli *et al.*, 2007). Il campione è composto da tre gruppi d'età (18-23; 24-29; 30-36 mesi), di 4 bambini ciascuno, registrati in due asili nido. La produzione verbale di ogni item del test è stata successivamente trascritta in simboli fonetici IPA e codificata al PC con programma freeware *Phon* (© 2006-2008 *The Phon Project*), versione 1.3R500, che è stato sviluppato all'interno della comunità scientifica CHILDES per informatizzare la ricerca sull'acquisizione fonologica.

La codifica in *Phon* del materiale raccolto è consistita nella creazione di un *record* per ogni parola realizzata dal bambino, codificata in caratteri alfabetici come glossa, direttamente in simboli IPA secondo la pronuncia del bambino e secondo la pronuncia adulta. In questo modo le forme delle due pronunce ricevono un allineamento automatico dei segmenti nei tipi sillabici costituenti la parola, che si basa sulle regole di sillabificazione dell'italiano. Nel presente studio verrà esemplificata la funzione relativa alle statistiche di frequenza ripartite per posizione del fono rispetto alla parola e alla sillaba.

2. INTRODUZIONE

Questo articolo ha come duplice obiettivo quello di descrivere un nuovo test, ancora in fase di sviluppo, per la raccolta e la valutazione delle capacità fonetico-fonologiche di bambini in un'età compresa tra i 18 e i 36 mesi, e un nuovo programma, *Phon*, per la codifica e l'analisi semiautomatica degli aspetti segmentali del parlato. Il nesso tra i due oggetti sarà qui dato dall'applicazione di *Phon* ai primi dati raccolti su un gruppo ristretto di bambini con sviluppo di linguaggio tipico, tramite una prima versione del test fonetico costruita alla fine dell'anno 2007. L'analisi dei dati raccolti con questo studio pilota ha permesso di rivedere alcuni punti del test, e di metter così a punto una seconda versione di cui faremo cenno nella discussione.

Nel campo delle ricerche sullo sviluppo fonetico, l'aspetto più studiato è senza dubbio quello segmentale, a causa di molteplici esigenze, sia di carattere pratico che teorico. Le esigenze di tipo pratico hanno a che fare con la difficoltà di usare strumentazioni e procedure non adatte a soggetti generalmente poco o per nulla collaborativi come i bambini, e con la necessità di valutare la normalità dello sviluppo con un metodo (apparentemente) facile e rapido quale quello basato sulla trascrizione fonetica del percetto uditivo. Nella pratica clinica, infatti, spesso è necessario confrontare la produzione segmentale di un certo soggetto con i dati normativi per la stessa fascia d'età, al fine di stabilire la normalità o meno di quel percorso individuale. L'attenzione agli aspetti segmentali, però, ha anche motivi teorici, quali la forte relazione di questo tipo di studi con un tipo di fonologia basata sul segmento e sui tratti distintivi.

Da un punto di vista teorico, raccogliere il parlato spontaneo infantile prodotto senza alcuna sollecitazione può risultare attraente per la sua ecologicità. Tuttavia, esistono rischi potenziali che possono minacciare la validità di questa procedura, soprattutto nel caso di bambini con problemi di linguaggio.

Secondo Shriberg & Kwiatkowski (1985), i bambini potrebbero:

1. essere restii a comunicare all'interno del tempo disponibile;
2. comunicare in modo inintellegibile;
3. produrre una quantità di parlato che differisce strutturalmente dai dati normativi, per quanto riguarda la distribuzione proporzionale di forme grammaticali, forme lessicali e fonemi;
4. reagire alla situazione di campionamento parlando in modo innaturale (a bassa voce, in modo canzonatorio o recitante, o canticchiando cfr. *playful speech register*, Shriberg & Kwiatkowski,1985).

Le componenti della procedura di campionamento che generano i problemi citati sopra riguardano il ruolo dell'esaminatore e il tipo di materiale usato per elicitare il parlato. Secondo Shriberg & Kwiatkowski (1985), le varie combinazioni di tipologie del ruolo dell'esaminatore e del materiale usato portano a identificare 5 condizioni di elicitazione (v. tab. 1). Questi autori escludono il parlato su ripetizione immediata, in quanto favorirebbe l'accuratezza delle produzioni del bambino, e dunque sovrastimerebbe la sua competenza fonologica (cfr. Kresheck & Socolofsky, 1972); per una riconsiderazione di un campione di parlato su ripetizione immediata, soprattutto quando il controllo sui vari fattori richiede l'uso di non-parole, si può consultare Edwards & Beckman (2008).

Queste cinque tipologie di raccolta di parlato infantile sono state testate da Shriberg & Kwiatkowski (1985), per l'efficacia da loro raggiunta in relazione alla produttività, all'intelligibilità, alla rappresentatività e alla reattività del bambino agli stimoli. Dalle loro statistiche risulta che, sebbene non ci siano grandissime differenze tra i risultati ottenuti con le diverse procedure, l'ultima tipologia di campionamento, cioè la denominazione di una lista di parole precostituita (o test di articolazione), ottiene sempre i migliori punteggi eccetto che per l'aspetto della produttività. La nostra attenzione si concentrerà quindi su questo tipo di procedura.

tipo di raccolta del campione	*tipo di parlato connesso*	*controllo sul contenuto*	*selezione e tipologia dei materiali di stimolo*	*Commenti e sollecitazioni dell'esaminatore*
1. Libero	Non contestualizzato (*directed*)	Non controllato	Materiali ed argomenti selezionati dal bambino	Limitato a commenti non direttivi
2. Storia	Non contestualizzato	Indiretto	Materiali ed argomenti selezionati dall'esaminatore: casetta *colorform*, pupazzi o marionette, argomenti scelti dal bambino sui materiali	Limitato a commenti non direttivi relativi agli stimoli
3. Routines (scenari)	Non contestualizzato e contestualizzato	Indiretto e diretto	Materiali ed argomenti selezionati dall'esaminatore: casetta *colorform*, che rappresenta tutte le consonanti da produrre. Gli argomenti possono essere più o meno correlati al materiale	Uso di domande e commenti per stimolare la verbalizzazione sui materiali
4. Intervista	Non contestualizzato e contestualizzato	Diretto	Nessun materiale. Gli argomenti possono essere scelti dall'esaminatore o dal bambino	Domande per identificare e proseguire gli argomenti su cui il bambino parla di più
5. Denominazione/lista (*scripted*)	Contestualizzato	Diretto	Materiali scelti dal'esaminatore, immagini di un libro che evocano parole ed argomenti che rappresentano tutte le consonanti da produrre	Domande relative alle immagini

Tabella 1: Variabili presenti nelle procedure dei test fonetici di raccolta e valutazione del linguaggio infantile (traduzione da Shriberg & Kwiatkowski, 1985, ad opera di chi scrive)

Mentre nel mondo anglosassone esiste un buon numero di test di articolazione per la valutazione delle capacità fonetiche e fonologiche dei bambini piccoli, in genere a partire dai 18-24 mesi, età in cui si raggiunge e consolida il cosiddetto Primo Vocabolario (cfr. per es. la rassegna di Shriberg & Kwiatkowski, 1985), per l'italiano l'unico test di una certa solidità scientifica e di buona diffusione è quello delle 'Prove per la Valutazione Fonologica del Linguaggio Infantile' (PFLI) di Bortolini (1995). Nelle parole di Bortolini (1995: 3), tali prove "sono state ideate per l'analisi clinica dei bambini con disordine fonologico […] sono tuttavia utilizzabili con tutti i tipi di linguaggio infantile" dai 2 ai 5 anni, e si compongono di un set di 90 figure [suddivise in un gruppo più piccolo di schede verdi, necessarie e sufficienti per la valutazione, e schede rosse, per eventuali approfondi-

menti e integrazioni, N.d.A.] per la raccolta del campione di linguaggio e di 15 schede per la valutazione".

Le procedure sono essenzialmente tre:

1. la descrizione delle capacità fonetiche del bambino, indipendentemente dal loro rapporto con il linguaggio adulto. Questa descrizione comporta l'inventario delle consonanti, delle vocali e dei gruppi consonantici nelle diverse posizioni della parola;
2. l'analisi contrastiva e la valutazione del sistema fonologico;
3. l'analisi in processi e la 'valutazione evolutiva'.

A nostra conoscenza questo strumento ha ricevuto solo due recensioni, in verità molto sintetiche. Doimo (1998: 153) sottolinea l'importanza di utilizzare un approccio all'analisi del sistema fonologico del bambino (e ai suoi disturbi) all'interno di un sistema di valutazione più ampio che enfatizza l'organizzazione linguistica generale e la variabilità fonetica interindividuale. Doimo (1998) riconosce che il test non è stato sottoposto a standardizzazione, ma il suo giudizio è tutto sommato positivo, senza però riportare alcun dato sperimentale a sostegno della sua valutazione. Rosolen & Barca (1999) invece si concentrano sulla criticità dell'applicazione del test riportandone i risultati su un campione sperimentale di 14 soggetti di controllo e 3 soggetti con DSL. Le autrici sottolineano come a fronte di un tempo medio di amministrazione di circa 45 minuti per entrambe le tipologie dei soggetti (Doimo considera come tempo di somministrazione 30-60 minuti), quello di trascrizione sia di ben 150 minuti. Inoltre, solo un soggetto di controllo e due con DSL riescono a produrre più di 100 parole (campione minimo richiesto) con il primo gruppo di schede (verdi) e solo 4 soggetti, (2 del gruppo di controllo e due DSL) superano le 200 parole totali (campione ottimale). In particolare Rosolen & Barca (1999: 219) non riuscirono a valutare i bambini DSL in quanto "i dati del manuale della prova sono insufficienti". Altre criticità rilevate dagli autori riguardano lo scarso potenziale di elicitazione di ben 10 schede e il fatto che ci siano fonemi poco rappresentati.

Nella nostra esperienza, e alla luce della trattazione di Shriberg & Kwiatkowski (1985), possiamo aggiungere questi altri punti deboli: assenza di standardizzazione, inconsistenza dell'aspetto psicometrico (la *performance* del bambino non riceve un punteggio), inconsistenza scientifica dei riferimenti normativi (Bortolini 1995, non spiega come è stato raccolto e in cosa consiste il campione su cui ha costruito le cronologie di riferimento pseudo-normativo), e tutti gli svantaggi esposti sopra che sono relativi alla sollecitazione di linguaggio connesso (vs denominazione), che qui puntualizziamo per il caso in questione:

1. produzione non controllata per numero di occasioni offerte a ciascun fono di essere prodotto (non solo alcuni foni sono poco rappresentati, ma altri lo sono anche troppo);
2. non confrontabilità interindividuale e intraindividuale;
3. tutto il materiale deve essere trascritto in IPA: enorme dispendio temporale;
4. il materiale di tipo iconografico come quello delle schede di Bortolini (1995); è poco adatto a bambini piccoli (almeno fino a 30 mesi);
5. troppo dipendente dalla loquacità (*talkativeness*);
6. potenziale difficoltà da parte dell'operatore ad individuare i target prodotti da un bambino poco intellegibile.

Il test presentato in questo articolo denominato 'Test Fonetico della Prima Infanzia' (TFPI) riguarda la valutazione delle capacità articolatorie in bambini di età compresa tra i 18 e i 36 mesi. È stato progettato da Serena Bonifacio e Claudio Zmarich nell'ambito di un progetto di ricerca corrente dell''IRCCS Burlo Garofolo' di Trieste. Il TFPI si propone la stesura dell'inventario fonetico, basato sulla produzione verbale del bambino (stimolata ma, se possibile, non ripetuta), per accertare se un fono o un gruppo consonantico risultano acquisiti in sede iniziale o non iniziale di parola (cfr. Zmarich & Bonifacio, 2005).

Gli item lessicali sono stati scelti in base alle esigenze di rendere il test veloce da somministrare e da analizzare tenendo conto dei tempi attentivi che un bambino di 18 mesi può dedicare alla prova (che possono essere anche inferiori a 30 minuti).

All'interno di queste esigenze, i criteri che hanno guidato le nostre scelte sono i seguenti:

1. Criterio Fonetico: i fonemi consonantici della lingua italiana devono essere attestati in almeno tre parole diverse per ciascuna posizione lessicale, quali: singola iniziale di parola, in gruppo consonantico iniziale di parola, singola intervocalica, in gruppo consonantico intervocalico, e tenendo in debito conto le restrizioni fonotattiche (cfr. Muljacic, 1972). Il gruppo consonantico iniziale è sempre omosillabico, il gruppo consonantico intervocalico può essere anche eterosillabico. Le tre parole, intese come possibilità offerte al bambino, devono garantire la produzione di almeno 2 token (repliche) del fono elicitato, venendo così incontro al criterio minimo di attestazione di un determinato fono nella procedura di compilazione degli Inventari Fonetici proposta da Stoel-Gammon (1985; cfr. anche Zmarich & Bonifacio, 2005).
2. Criterio Semantico/Frequenziale: le parole utilizzate nel test, tratte dall'appendice A del libro Parole e Frasi 'Primo Vocabolario del Bambino' di Caselli, Pasqualetti & Stefanini (2007), sono state selezionate tra le parole della categoria dei sostantivi (criterio semantico) in base al valore percentuale più alto all'interno di ciascuna delle 3 fasce di età (criterio frequenziale). Ricordiamo che i valori percentuali si riferiscono alla proporzione di bambini sul campione totale esaminato da Caselli *et al.* (2007), i cui genitori attestano che sono in grado di produrre la parola in questione. I sostantivi devono essere concreti, per poter essere rappresentati tramite oggetti o figure.
3. Criterio della gradualità nella Complessità Fonetica: le parole utilizzate mostrano una progressiva complessità per numero e tipi di sillabe (cioè con gruppi consonantici via via più complessi e/o più rari), così come queste emergono dagli studi di Zmarich e Bonifacio (2005), Zmarich, Dispaldro, Rinaldi & Caselli (2009) e Zmarich, Dispaldro, Rinaldi & Caselli (accettato per la pubblicazione). Infatti, come osservato anche da Edwards & Beckman (2008), raramente nei test di articolazione la lunghezza della parola (determinata in base al numero delle sillabe) viene presa esplicitamente in considerazione, e si può riscontrare che due diverse consonanti vengano proposte in parole di lunghezza diversa. Le conseguenze possono essere gravi poiché dallo studio di Edwards & Beckman (2008) risulta che i bambini sono significativamente sensibili a questo fattore, a tal punto che la mancata pronuncia di una certa consonante potrebbe essere dovuta più alla lunghezza della parola che la contiene che all'effettiva incapacità del bambino a produrla.

Per quanto riguarda numero e tipi di sillabe, le parole della fascia 18-23 mesi sono soprattutto bisillabiche, con pochi nessi consonantici omosillabici, esclusivamente del tipo consonante + /j/ o /w/; quelle della fascia 24-29 mesi sono anch'esse in prevalenza bisillabiche, ma aumenta la proporzione relativa di parole trisillabiche, e di nessi biconsonantici di tipo eterosillabico e omosillabico, che non sono più limitati alla tipologia C + /j/ o /w/. Infine, nella versione del test che è stato amministrato ai bambini dello studio pilota, le parole della fascia 30-36 mesi sono distribuite in modo che la percentuale di parole bisillabiche sia del 44,4% e quello delle parole trisillabiche 30,8%, mentre le quadrisillabiche sono maggiormente presenti rispetto alle fasce precedenti ma in percentuale minore rispetto alle altre due strutture. I nessi consonantici di tipo omosillabico sono abbastanza frequenti, con qualche nesso triconsonantico.

Un altro fattore considerato importante nella scelta degli item lessicali da proporre al bambino è per Edwards & Beckman (2008) la fonotassi, cioè la probabilità statistica che un fono ha nel comparire insieme a un altro, in questo caso la vocale che segue la consonante da elicitare. All'interno di questo fattore va considerata anche la prominenza accentuale (poiché i bambini producono meglio le consonanti delle sillabe accentate). Come hanno dimostrato questi autori, le consonanti che nei test di articolazione sono seguite da vocali con cui nel parlato si accoppiano raramente, sono significativamente meno prodotte dai bambini rispetto alle stesse consonanti seguite da vocali che ne parlato le accompagnano più frequentemente. La scelta di Edwards & Beckman (2008), nel formare un *corpus* da far produrre ai bambini, è stata quella di farli produrre su ripetizione immediata di non-parole, ma questa scelta è stata imposta anche e soprattutto dalla loro prospettiva di tipo cross-linguistico (le restrizioni fonotattiche sono diverse da lingua a lingua). Nel TFPI è stato fatto uno sforzo generico per evitare i contesti vocalici più rari, e nel mantenere uno schema accentuale di tipo parossitono (parole accentate sulla penultima sillaba), ma questo obiettivo non è stato perseguito in modo esplicito e sistematico.

Nella procedura di somministrazione del TFPI, la produzione verbale dell'item lessicale viene sollecitata presentando al bambino l'oggetto-giocattolo, rappresentante il target della lista prevista per la sua fascia d'età.

Ciascuno stimolo viene presentato singolarmente, chiedendo al bambino: "cos'è? come si chiama questo?" non più di tre volte. Se il bambino produce il nome immediatamente dopo l'ostensione dell'oggetto e la richiesta dell'operatore, si considera questa produzione come prodotta spontaneamente. Se invece il bambino non dice il nome neanche al terzo tentativo, l'operatore denomina l'oggetto indicandolo al bambino, e chiedendogli di ripetere la parola, in questo modo: "questo si chiama X. Come si chiama questo"? Se il bambino finalmente produce la parole, questa viene considerata come ripetizione immediata. Se non la produce non viene più ripresentata. I due tipi di produzione, spontanea o ripetuta, ovviamente riceveranno punteggi diversificati a favore della produzione spontanea, ma al momento quest'aspetto non è stato risolto in modo soddisfacente.

Un aspetto peculiare di questo test è la sua tripartizione in liste lessicali diverse da somministrare in base all'età del bambino. Questo è stato fatto per ridurre il tempo di coinvolgimento del bambino, che deve essere tanto minore quanto più il bambino è piccolo, e per poter testare i bambini di età diverse sui diversi aspetti che sono legati alla competenza fonetico-fonologica. Tra questi aspetti c'è la capacità di pronunciare un fono isolato o in gruppo consonantico (e in gruppo omosillabico o eterosillabico), che risiede in posizione iniziale o intervocalica, in parole corte o lunghe per numero di sillabe. Come

abbiamo segnalato prima, nel costruire ciascuna lista, che combina in modo diverso questi fattori così da risultare graduale, abbiamo tenuto conto della frequenza delle parole nelle tre fasce di età (Caselli *et al.*, 2007), degli inventari fonetici prodotti dai soggetti normali dai 18 ai 30 mesi in un *setting* semistrutturato simile agli scenari elencati al n. 3 della tabella n.1 (Zmarich & Bonifacio, 2005), e delle caratteristiche fonetiche delle parole del PVB analizzate in Zmarich *et al.* (2009) e Zmarich *et al.* (accettato per la pubblicazione).

Questa interruzione di continuità dai 18 ai 36 mesi però potrebbe potenzialmente creare un artefatto o pregiudizio nella comparazione dei bambini attraverso le fasce di età. Prendiamo il caso, per esempio, di un bambino della seconda o della terza fascia di età che non riesce a produrre una certa consonante. Per escludere che per questo bambino l'ostacolo sia rappresentato dalla lunghezza di parola più che dalla consonante o gruppo consonantico in sé e per sé, abbiamo inserito quelle consonanti che in base a Zmarich e Bonifacio (2005) sono tipicamente acquisite in una fascia successiva, in una parola di lunghezza (per n. di sillabe) e complessità (per gruppo consonantico) simile a quelle della fascia precedente (dunque relativamente corte e poco complesse), mentre solo le consonanti già attestate nella fascia di età precedente sono state usate in parole relativamente lunghe e con gruppi consonantici relativamente complessi.

Un altro potenziale problema è rappresentato dai bambini eccezionalmente dotati, e dai bambini ai confini di fascia (per esempio, di 23 mesi per la prima fascia, di 29 mesi per la seconda fascia ecc..). Questi soggetti potrebbero produrre correttamente tutte le parole richieste per la loro fascia, e ancora essere diversi tra loro (per competenza), senza che questa diversità possa emergere a causa del raggiunto tetto. Allo scopo di valutarli e classificarli correttamente, è stata prevista la possibilità che un bambino che produca in modo accurato tutte, o quasi tutte, le parole di una certa lista (la soglia per poter passare alla lista successiva si potrebbe fissare alla produzione del 90% delle parole), possa accedere alle parole della fascia successiva che contengono le strutture non ancora prodotte (siano essi foni isolati e in gruppo consonantico, o parole lunghe per numero di sillabe).

Qui di seguito riportiamo un elenco di caratteristiche che un test di articolazione ideale dovrebbe possedere, tratte da tre articoli di Smit e collaboratori in cui sono presentati i criteri di costruzione e validazione di uno tra i più famosi ed usati test di articolazione negli USA, lo *Iowa-Nebraska Articulation Test* (cfr. Smit, Hand, Freilinger, Bernthal & Byrd, 1990; Smit, 1993a; Smit, 1993b).

<u>Soggetti</u>:

- In fase di validazione e standardizzazione di un test, bisogna includere anche i risultati tratti da soggetti con ritardo di linguaggio e disordini fonologici, nella proporzione prevista dall'incidenza delle patologie del linguaggio e della parola nelle fasce d'età previste. Ad es., a 24 mesi circa il 15% dei bambini presenta una dimensione del vocabolario espressivo =/< a 50 parole (Desmarais, Sylvestre, Meyer, Bairati & Rouleau, 2008);
- in fase di validazione e standardizzazione di un test, bisogna tener conto di variabili demografiche come il genere, la densità della popolazione, il livello di istruzione, la lingua parlata in famiglia (anche se in Smit *et al.*, 1990, viene poi dimostrato che di tutte le variabili socioeconomiche l'unica che si rivela significativamente importante è il genere dei soggetti, maschile o femminile, e non a tutte le età);

- alcuni soggetti, soprattutto nella prima fascia, possono non avere completato la dentizione, con conseguenti problemi nella produzione dei suoni dentali e labiodentali;
- l'età dei bambini deve andare dai 2 anni ai 9 anni, a intervalli di 6 mesi.

Aspetti del test:

- Nella scelta del materiale fotografico (o degli oggetti) bisogna lasciar decidere a un gruppo di bambini il miglior rappresentante di ogni tipo (a maggioranza);
- la variabilità nei punteggi diminuisce in base all'età.

Aspetti della trascrizione:

- Per la costruzione delle curve di normalità (percentili) per l'acquisizione di ogni fono o gruppo consonantico, bisogna allenare un gruppo di trascrittori, magari con esempi di trascrizione da video che vengono loro forniti, perché questi trascrittori dovranno poi trascrivere al meglio il campione raccolto che diventerà quello di riferimento.

Definizione del range di accettabilità per le risposte del bambino e precompilazione dei tipi più comuni di errore:

- Stabilire quali varianti accettare come realizzazione normale per ogni fono, e quali no (l'esempio italiano potrebbe riguardare alcune realizzazioni allofoniche di /r/);
- per le variabili da considerare errori, è meglio fornire nella versione finale (commerciale) del test la trascrizione di un certo numero di errori, in modo che il futuro somministratore possa ridurre i tempi e i rischi di errore nell'applicazione del test in ambito clinico, semplicemente contrassegnando la casella appropriata. È meglio però prevedere anche la possibilità di trascrivere *ex-novo* gli errori meno comuni.

Punteggio:

- Attribuzione di un punteggio di 1.0 potenziale per ogni singolo fono o *cluster* per ogni posizione. Se un fono è elicitato nella stessa posizione X volte, allora il punteggio diventa 1.0:X, per es. se un bambino produce solo una parola delle due proposte che comprendono il fono da produrre, allora il suo punteggio per quel fono sarà 0.5;
- il punteggio totale raggiunto dal bambino viene determinato sommando il punteggio degli item prodotti (v. punto precedente), esprimendolo in % sul punteggio totale dei producibili.

Standardizzazione:

- Per ogni fono vanno preparate delle curve di acquisizione basate sull'età;
- un fono può dirsi attestato quando è prodotto il 90% delle volte;
- un fono acquisito deve mostrare stabilità nel tempo (non deve abbassarsi sotto l'80% nell'arco di un anno);
- certi foni potrebbero non raggiungere mai il 100% di acquisizione (vedi /ŋ/ velare in Smit *et al.*, 1990).

Nessi consonantici:

- sebbene ogni singolo nesso è raro, le parole che contengono nessi sono abbastanza frequenti. I nessi sono importanti perchè bersaglio dei processi fonologici. L'incapacità di produrre nessi colpisce l'intelligibilità.

Fonetica vs Fonologia:

- Mentre per la fonetica basta accertare che il bambino sappia produrre un certo fono in una certa posizione (potenzialità espressiva) per la fonologia bisogna accertare la funzionalità della produzione ai fini della comunicazione, cioè quali sono i foni che contrastano per coppie minime (vedi per es. Dinnsen, 1992);
- importanza dei cosiddetti 'processi fonologici' (cfr. Ingram, 1981; per l'italiano, cfr. Bortolini, 1995), che rappresentano i vari modi sistematici con cui i bambini possono semplificare la struttura e il sistema fonologico delle parole adulte: se un bambino non produce un fono in una certa posizione, bisogna accertarsi che non sia per l'incapacità di pronunciare quel fono, ma per la presenza di un processo che colpisce tutti i foni in quella posizione (per es., nelle parole polisillabiche, il processo designato col nome di cancellazione della sillaba debole potrebbe impedire la produzione di qualsiasi consonante compaia in sillabe distanti, e per questo dette 'deboli', dalla sillaba accentata). È importante riportare la frequenza dei vari tipi di processi fonologici in base all'età, per poter poi classificare i bambini, in base ai processi fonologici esibiti, come normali, in ritardo, o devianti.

Altri requisiti che un test deve rispettare sono invece relativi agli aspetti di tipo psicometrico (cfr. Pedrabissi & Santinello, 1997). Vediamo per es. quali aspetti psicometrici sono considerati da Newcomer & Hammill (1997), per il test *TOLD P-3*:

- affidabilità di campionamento (la varianza dell'errore deve essere omogenea per tutti gli item in un test);
- affidabilità di costanza temporale (stessa varianza dell'errore nel test-retest);
- affidabilità di assegnazione dei punteggi (non bisogna lasciare troppa libertà e soggettività al valutatore);
- validità di contenuto (il test deve veramente impattare il tipo di comportamento che ci si propone di misurare: tipo di prove che propone; domande alla quali rispondere, ecc.);
- validità correlata al criterio (efficacia di un test a predire la performance individuale in attività specifiche: deve correlare bene con altri test che misurano le stesse attività);
- validità di costrutto: corrispondenza tra risultati ottenuti e costruzione teorica proposta.

Durante la fase sperimentale dell'applicazione del TFPI, è stato registrato un ampio campione di bambini tra i 18 e i 36 mesi, allo scopo di verificare la facilità di somministrazione, i tempi e soprattutto l'adeguatezza del materiale proposto, in base ai 3 criteri suesposti. La produzione verbale spontanea di 4 bambini (2 maschi e 2 femmine) di ciascuna fascia d'età è stata trascritta da trascrittori esperti di linguaggio infantile con l'Alfabeto Fonetico Internazionale (IPA, 1999). Per una serie di motivi, di tipo economico e

scientifico, che saranno evidenziati nel corso della descrizione di *Phon*, si è deciso di codificare tutte le realizzazioni infantili dei target, prodotte spontaneamente, e i target stessi, con questo programma.

Prima di presentare *Phon*, è necessaria però una breve premessa, che spieghi il contesto culturale in cui il programma è nato. Nella comunità internazionale degli studiosi dello sviluppo fonetico è molto sentita la necessità di studi interlinguistici sull'acquisizione fonetica: tali studi potrebbero infatti aiutarci a determinare se la struttura fonetica del babbling variato e del primo vocabolario (fino a 50 parole) è determinata principalmente dall'organizzazione di tipo universale del livello motorio, che si ritrova nel primo *babbling*, o se la selezione e l'organizzazione delle strutture fonetiche riveli un livello di elaborazione 'linguistico' di tipo linguospecifico, e, se si, quale (Boysson-Bardies de, Vihman, Roug-Hellichius, Durand, Landberg & Arao (1992).

Il presupposto di questi studi interlinguistici è la creazione di grandi database interlinguistici. I database interlinguistici possono essere creati solo attraverso l'uso di programmi informatici dedicati, possibilmente liberi, robusti, flessibili, relativamente semplici da usare, ad opera di una comunità di ricercatori che si dia delle regole per perseguire scopi di interesse comune. Il primo programma dedicato all'analisi degli aspetti segmentali del parlato infantile nasce all'interno della comunità CHILDES (http://childes.psy.cmu.edu/), quando molti dei ricercatori interessati agli aspetti fonologici di tipo segmentale dell'acquisizione linguistica che erano interessati a verificare l'universalità delle loro ipotesi teoriche, avendo bisogno di disporre di database interlinguistici, si trovarono limitati dalle funzionalità dei programmi fino ad allora sviluppati, CHAT (per la codifica) e CLAN (per l'analisi), che erano esclusivamente riservati agli aspetti conversazionali, sintattici, morfologici e lessicali.

Nonostante i tentativi fatti allo scopo di adattare CHAT e CLAN, si decise di creare un programma ex-novo, utilizzando però *software* proprietari. Il *ChildPhon* originale, utilizzato per alcuni studi sull'acquisizione della fonologia in olandese tra la fine degli anni '80 e i primi anni '90 all'Istituto Max Planck di Psicolinguistica da Paula Fikkert e Clara Levelt, rappresenta il primo tentativo di creazione di database interlinguistici. Tale programma, utile per l'archiviazione di forme trascritte foneticamente e di altre informazioni rilevanti su queste forme come il nome e l'età del bimbo, il tipo di produzione ecc., risentiva di un certo numero di svantaggi, il principale dei quali consisteva nell'impossibilità di incorporare dati multimediali. Attualmente molti studi sono basati proprio su questi tipi di dati. In secondo luogo, il database non eseguiva in modo automatico nessuna codifica o analisi. Infine, il programma non fu mai messo a disposizione della comunità linguistica più ampia. Il database, per le sue limitazioni tecniche ed i costi molto elevati, non è stato più utilizzato né sviluppato.

Si progettò quindi, tra il 1999 e il 2003, un *ChildPhon* di seconda generazione. Esso era dotato di integrazione audio digitale, pertanto offriva accesso facilitato ai campioni di parlato e permetteva di abbandonare le ingombranti trascrizioni su cassette. Questa nuova versione forniva anche una funzione di sillabificazione automatica e permetteva di comparare le forme effettivamente prodotte dal bambino al target. Questo software offriva alcuni tipi di analisi automatizzata e più vantaggi rispetto all'originale, ma le limitazioni relative all'uso di *FileMaker Pro* impedirono lo sviluppo di uno strumento più flessibile per la gestione del *datum* multimediale. Inoltre, le sue analisi automatiche erano troppo limitate

per coprire un ampio ventaglio di ricerche linguistiche. Pertanto, nel 2004, il progetto *ChildPhon* venne abbandonato a favore del progetto *Phon*.

Questo progetto partì nel 2003 con dei lavori di tesi all'interno della Carnegie Mellon University, e proseguì in stretta collaborazione con il team di programmatori (Byrne, Wareham e Rose) di B. MacWhinney che avevano già sviluppato altri software per il consorzio CHILDES.

Phon mira a fornire allo studio dello sviluppo fonologico un database che sia il più possibile flessibile, potente e funzionale. Com'è riportato sul sito di *Phon* (http://phon.ling.mun.ca/oldwiki/Phon/Manual), questo programma *software* facilita straordinariamente un gran numero di funzioni richieste per le analisi dello sviluppo fonologico. Per es., *Phon* supporta la connessione ai dati multimediali, la segmentazione delle unità, la trascrizione in doppio cieco, l'etichettatura automatica dei dati e la comparazione sistematica tra le forme fonologiche del target (il modello) e l'effettiva produzione del soggetto. Il programma lavora sia su *Mac OS X* sia su *Windows* ed è interamente compatibile con il formato CHILDES (cfr. CHAT e CLAN).

Phon è disponibile gratuitamente come *software open-source*, soddisfacendo i bisogni specifici di chiunque studi il linguaggio dal punto di vista segmentale (cioè quello basato sulla trascrizione fonetica), e debba comparare le realizzazioni fonetiche con i loro target, in particolare per campi quali:

- lo sviluppo fonologico di L1;
- l'acquisizione di L2;
- i disturbi di linguaggio.

Tutte queste funzioni sono accessibili attraverso un'interfaccia grafica *user-friendly*. Inoltre, i database codificati e analizzati in *Phon* possono anche essere interrogati utilizzando una interfaccia di ricerca potente. Si possono poi eseguire delle analisi di tipo avanzato e applicarle a una selezione particolare dell'intero campione in base alla presenza/assenza di determinate caratteristiche, utilizzando un linguaggio dalla sintassi abbastanza semplice, chiamato *PhonEx*. Questo linguaggio si serve di operatori logici su elementi definiti in termini di teoria linguistica: confini, segmenti, tratti distintivi. Le analisi possono così essere applicate a tutti i records, frasi o parole, del *corpus* contenuto nel file.

3. PROCEDURA SPERIMENTALE

3.1. Soggetti e setting di registrazione

Il Test Fonetico della Prima Infanzia (TFPI) è stato somministrato a bambini di età compresa tra 18 e 36 mesi reclutati in tre asili nido di Trieste, le cui famiglie avevano firmato il consenso alla partecipazione allo studio.

I criteri d'inclusione nello studio erano che i bambini fossero monolingui, nati a termine, non gemelli e con uno sviluppo di linguaggio espressivo rilevato con il PVB (Caselli, Pasqualetti & Stefanini, 2007) > al 5° percentile.

La prima fase di raccolta si è concentrata sulla lista delle parole target selezionata a priori sulla base dei tre criteri esposti nell'introduzione, per le tre fasce d'età 18-23, 24-29 e 30-36 mesi, allo scopo di verificare la facilità di somministrazione, tempi e adeguatezza del materiale proposto. L'intera produzione del bambino è stata audioregistrata con strumenti

professionali quali registratori digitali *hi-fi* (*Edirol R-09*) e microfoni direzionali accuratamente posizionati.

Al termine della fase sperimentale di raccolta si è deciso di selezionare 12 soggetti che avevano esibito un buon livello di motivazione e collaborazione al test, manifestando un buon grado di intelligibilità dell'eloquio.

Seguendo la metodologia proposta da Stoel-Gammon (1985) per la compilazione degli inventari fonetici, si è deciso di non considerare le parole che mostravano una scarsa somiglianza fonetica con la forma adulta, cioè che contenevano meno di due suoni consonantici in comune con quest'ultima.

3.2. Codifica con Phon

Si è proceduto quindi alla creazione di un record per ciascuna parola realizzata dal bambino (vedi Fig. 1, tratta dal manuale di PHON), che è stata codificata sia in caratteri alfabetici, come glossa, nel livello d'informazione *Ortography*, sia in simboli IPA, nel livello d'informazione *IPA actual* per la pronuncia del bambino, e nel livello d'informazione *IPA target* per la pronuncia adulta. Le produzioni ottenute su ripetizione immediata dell'item lessicale sono state contrassegnate inserendo la dicitura 'parola ripetuta' nelle note presenti in basso di ciascun record, per poi consentire un eventuale scorporo delle produzioni ripetute da quelle spontanee.

Le due produzioni, il target lessicale e la realizzazione infantile del target hanno ricevuto in modo automatico una suddivisione in sillabe (*target syllables*, *actual syllables*) e un allineamento delle due trascrizioni (*syllable alignement*), in base alle regole di sillabificazione dell'italiano. Qui va sottolineato che *Phon*, nel fare l'allineamento, attribuisce in modo automatico i segmenti di una sillaba a Onset, Nucleo e Coda, distinguendoli con colori diversi, in base alla loro posizione, alla lingua e al modello teorico adottato. Per l'italiano, queste regole sono state implementate da Y. Rose secondo l'indicazione di C. Zmarich, che ha seguito prevalentemente la trattazione di Bertinetto & Loporcaro (2005).

In base a queste regole, i punti tradizionalmente più problematici della sillabificazione dell'italiano sono stati risolti nel modo seguente: il fono [s] in posizione iniziale di parola in nesso consonantico è stato considerato appendice sinistra, mentre in posizione iniziale di nesso consonantico intervocalico è stato considerato come coda della sillaba precedente (es.: 'as.ta' e non 'a.sta'). Per quanto riguarda i foni approssimanti [j w], sono stati codificati come consonanti e più precisamente considerati come onset (o attacco) della sillaba di cui fanno parte, se seguiti da vocale, o come coda di sillaba, se preceduti da vocale. Ricordiamo inoltre che ogni record è stato associato in modo definitivo alla porzione di segnale acustico che lo riguarda, in modo tale che in qualsiasi momento è possibile verificare la trascrizione ascoltando l'audio corrispondente.

Phon consente diversi tipi di analisi di uso più comune nello studio dell'acquisizione fonologica. È possibile applicarli dalla finestra del *Session Manager* ai *records* della singola sessione. In questo caso col menù *Inventory* si possono ottenere subito le statistiche di frequenza dei foni, dei tipi sillabici oppure dei pattern accentuali sia del target che della produzione effettiva.

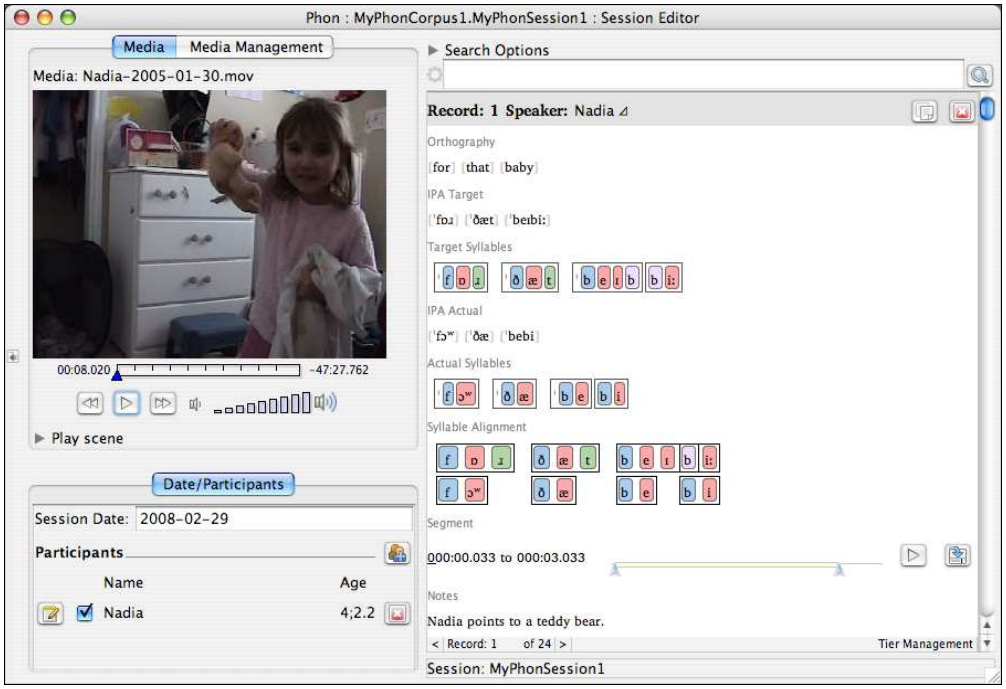


Figura 1: Finestra della Session Editor (dal manuale di Phon)

È possibile inoltre ottenere l'analisi di molti tipi di processi fonologici (cfr. Ingram, 1981; Bortolini, 1995), che semplificano la struttura e il sistema fonologico delle parole adulte, e che anche grazie all'allineamento sillabico trovano una facile applicazione analitica, ricorrendo al menù *Search* della finestra del *Project Manager*. Per questo articolo si è deciso di applicare le analisi di frequenza dei foni e dei tipi sillabici alle configurazioni fonetiche dei target e delle produzioni effettive dei bambini, al fine di operare, tra le due categorie, confronti che fossero significativi delle reali capacità dei soggetti. Utilizzando la funzione del menu *Inventory*, abbiamo ottenuto un conteggio della frequenza di ciascun fono, sia consonantico che vocalico, all'interno di ciascun soggetto. Questi conteggi sono espressi come una matrice le cui righe sono costituite dai record (cioè le 'parole') e le colonne dai simboli fonetici. Per ogni record, la riga elenca una successione di cifre che indicano il numero di volte in cui il fono relativo è presente (per es. nel record 'mami' avremo in corrispondenza di /m/ il n. 2, e in corripondenza di /a/ e /i/ il numero 1, mentre tutti gli altri foni riporteranno il valore 0). L'ultima riga della matrice contiene il totale delle frequenze dei foni in tutti i record del soggetto. In questa sede sono stati presi in considerazioni solo i foni consonantici, per ciascuno dei 12 soggetti. Questi sono stati codificati in simboli fonetici SAMPA (http://www.phon.ucl.uk/home/sampa/) ed inseriti in una matrice di *Systat 10.0* per l'analisi statistica. Per effettuare un'analisi sui risultati di gruppo si è proceduto alla creazione di tre fasce, ciascuna delle quali costituita dalla somma delle frequenze totali di ciascuno dei foni consonantici per i quattro soggetti di ogni fascia, in modo che la Fascia 1 raggruppasse quelli dei 18-23 mesi, la Fascia 2 dei 24-29 mesi, la Fascia 3 dei 30-36 mesi. Infine sono stati calcolati i valori percentuali per fascia, rapportando la somma delle frequenze di ciascun fono nei 4 soggetti al totale di tutti i foni

prodotti. Questo calcolo è stato eseguito prima sulle configurazioni fonetiche dei target, per accertarci che tutti i foni fossero stati proposti al bambino che li doveva produrre, e poi sulle produzioni effettive.

4. RISULTATI

4.1 Conteggi di frequenza dei foni consonantici

La figura 2 illustra la frequenza percentuale delle consonanti del target elicitati. È immediatamente evidente come le consonanti delle parole proposte al bambino abbiano frequenze diverse (in base al fatto che molte consonanti non ricorrono solo nelle parole previste per loro, ma anche nelle parole previste per l'elicitazione di altre consonanti). Risulta poi altrettanto evidente come ci sia un aumento progressivo dei valori percentuali delle consonanti affricate e della vibrante /r/, che rispecchiano l'aumento della complessità fonetica delle parole proposte al bambino. Le consonanti maggiormente proposte sono le occlusive e le nasali, presenti soprattutto nella fascia 1, e la laterale /l/. La frequenza delle consonanti prodotte dai bambini è esposta in figura 3. Nella prima fascia c'è una grande preponderanza delle consonanti occlusive, soprattutto sorde, che sono molte di più di quelle proposte nei target, probabilmente per la presenza di processi fonologici come ad es. lo *stopping*, e di sostituzioni di foni non ancora in repertorio, con foni già in repertorio, come l'intera serie delle consonanti occlusive. È da notare che le consonanti sonore sono poche ma tendono ad aumentare con l'età. Si può poi osservare la produzione sporadica di foni estranei al sistema fonologico della lingua italiana, come [x] o [θ] (trascritto in SAMPA come 'T'), che rappresentano esiti di processi fonologici o sostituzioni di fonemi target.

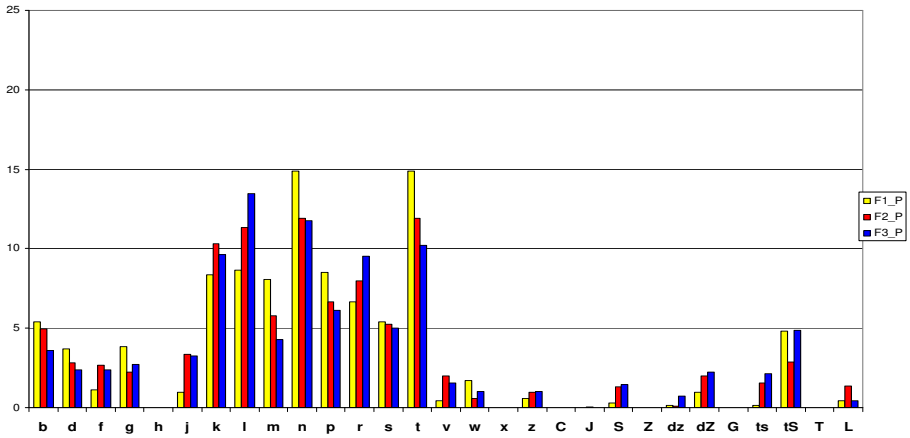


Figura 2: frequenza percentuale delle consonanti nei target elicitati sul totale dei foni prodotti per ogni fascia di età. Le consonanti sono scritte in alfabeto fonetico SAMPA e seguono un ordine pseudo-alfabetico

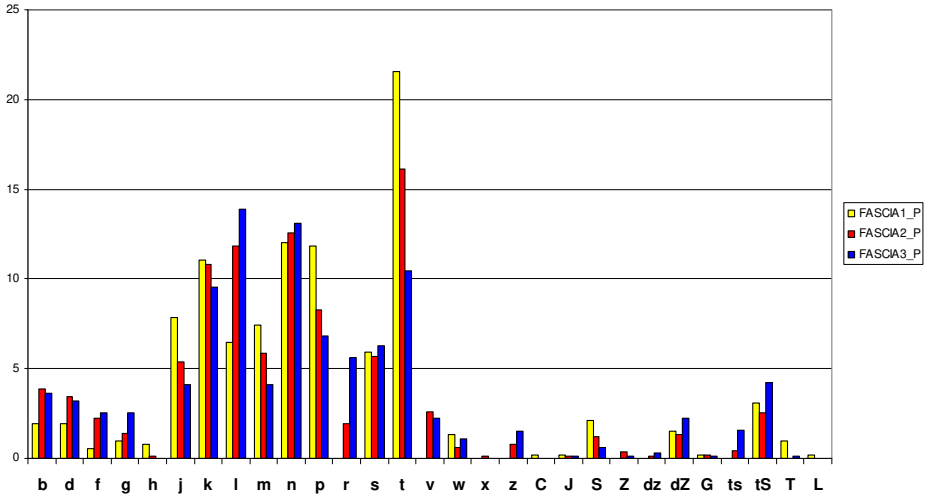


Figura 3: Frequenza percentuale delle consonanti nelle produzioni dei bambini sul totale dei foni prodotti per ogni fascia di età. Le consonanti sono scritte in alfabeto fonetico SAMPA e seguono un ordine pseudo-alfabetico

4.2 Conteggi di frequenza dei tipi sillabici

Passando a considerare le frequenze percentuali dei tipi sillabici relativi ai target elicitati, esposte in figura 4, si può osservare come siano state proposte parole con un'assoluta prevalenza dei tipi sillabici CV, che si situano appena al di sotto del 70%, e CVC, circa il 20%.

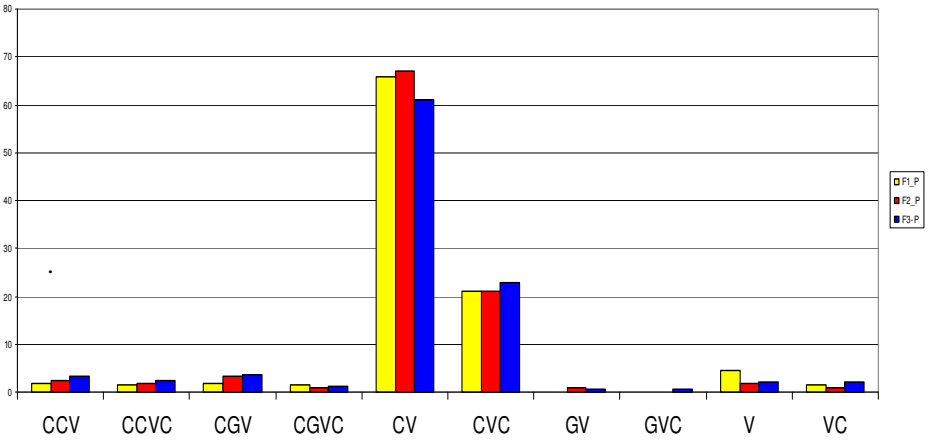


Fig. 4: Frequenza percentuale dei tipi sillabici dei target elicitati sul totale dei tipi sillabici prodotti per ogni fascia di età (G=Glide)

Inoltre, soprattutto per la fascia 1, sono state proposte parole che comprendono il tipo sillabico con sola vocale, che però diminuisce nei target proposti della seconda e terza fascia.

Le frequenze percentuali dei tipi sillabici effettivamente prodotti sono esposte in fig. 5. Si nota la tendenza, da parte dei bambini delle tre fasce, a produrre soprattutto parole con configurazione sillabica del tipo CV, che infatti con oltre il 70%, è superiore al valore percentuale che questo tipo sillabico ha nel target. Questo dato ci fa supporre che i bambini mettano in atto parecchie semplificazioni. Inoltre il tipo sillabico CVC si riduce di molto raggiungendo appena il 10%, mentre nel target è sempre sopra il 20%.

Nello studio dell'intero *corpus* sono state tralasciate le analisi dei processi fonologici messe in atto dai bambini. In futuro ci si propone di verificare quali parole di quelle utilizzate nel test fonetico sono state maggiormente soggette a semplificazione. Ciò è importante al fine di valutare l'adeguatezza dei target selezionati evitando così quelli che danno facilmente adito a processi fonologici, in cui si può assistere alla cancellazione del fono che si voleva elicitare nel bambino. Inoltre, sempre per verificare nella maniera più completa la validità del test, andranno considerate anche le varianti dialettali e le parole non prodotte, cercando di capire se la loro mancata produzione è dovuta alla mancata conoscenza da parte del bambino di quel target, o è dovuta alla inadeguatezza dello stimolo visivo (oggetto o figura) a rappresentare il referente di quella parola.

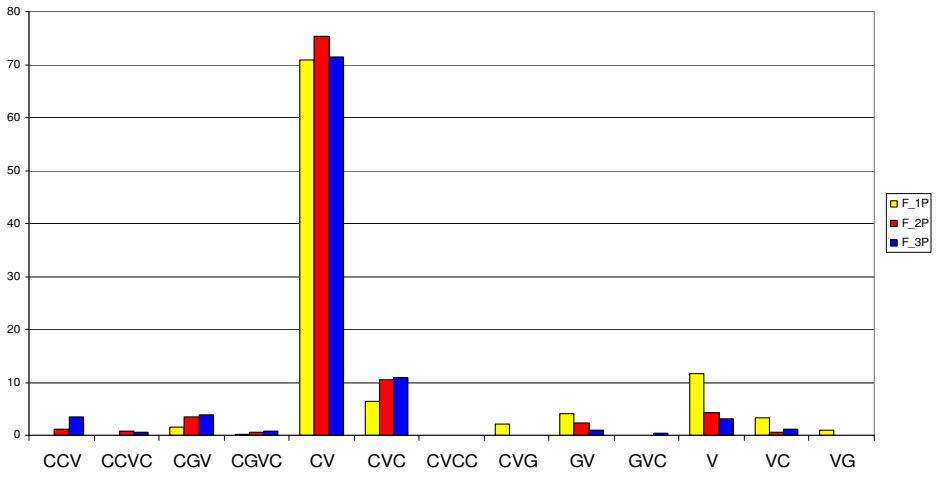


Figura 5: Frequenza percentuale dei tipi sillabici delle produzioni effettive dei bambini sul totale dei tipi sillabici prodotti per ogni fascia di età (G=Glide)

5. DISCUSSIONE

A conclusione di questo articolo, vorremmo fare due ordini di considerazioni, che riguardano il Test Fonetico della Prima Infanzia e il programma *Phon*.

Per quanto riguarda il test fonetico, presentato per la prima volta in questa sede, benchè sia ancora prematuro pronunciarsi con certezza sulla sua validità ed applicabilità in ambito

clinico, i dati raccolti finora promettono bene. È importante infatti esaminare, ad esempio, l'adeguatezza dei target lessicali, che sono stati scelti ed inseriti nel test, e degli stimoli visivi (oggetto o figura) utilizzati per elicitare la produzione del bambino, e i dati finora raccolti si sono rilevati utili per valutare questi aspetti. Una volta pervenuti ad un formato quasi definitivo, il passo successivo sarà quello di raccogliere dati su un campione che sia statisticamente rappresentativo per fornire dei profili normativi.

La versione del test utilizzata in questo studio risale alla fine del 2007, e presentava alcune criticità che in gran parte sono state corrette nella versione corrente, mentre per altre si devono ancora trovare soluzioni soddisfacenti. Senza entrare nel merito delle singole scelte, possiamo tuttavia descrivere i criteri che ci hanno portato a modificare profondamente la forma del test attuale.

Come prima osservazione, la lista di parole (vedi appendice 1) soprattutto per i 36 mesi era troppo lunga e ciò richiedeva tempi di somministrazione e di attenzione per i bambini, e di analisi dell'operatore, piuttosto lunghi. Inoltre, i bambini erano portati a semplificare molte parole tramite errori di sostituzione o i cosiddetti processi fonologici. Al momento non abbiamo usato *Phon* per l'analisi dei processi fonologici messi in atto dai bambini dell'intero *corpus*. Abbiamo condotto un'analisi più superficiale individuando quanti bambini avevamo prodotto la parola spontaneamente e/o su ripetizione, e nel caso l'avessero prodotta, se fosse corretta o fosse stata semplificata in modo tale da pregiudicare la produzione del fono per il quale la parola era stata proposta. Ciò è importante al fine di valutare l'adeguatezza dei target selezionati eliminando così quelli che non vengono prodotti frequentemente, quelli che danno facilmente adito a processi fonologici, che portano alla scomparsa del fono, quelli che hanno varianti dialettali, e quelli che possono dare adito a creazione di diminutivi o onomatopee. Nel caso della mancata produzione bisogna poi cercare di capire se questa è dovuta alla mancata conoscenza da parte del bambino di quel target, o è dovuta alla inadeguatezza dello stimolo visivo (oggetto o figura) come referente di quella parola.

I provvedimenti che abbiamo preso sono stati di tre tipi:

- Per i 36 mesi abbiamo previsto la possibilità di somministrare i target tramite presentazione di figure, e non più di oggetti. La loro origine e il loro formato si presta anche ad una presentazione tramite PC.
- La lista è stata drasticamente ridotta, eliminando la necessità che ogni fono in ciascuna posizione debba essere rappresentato da 3 parole diverse, per poter essere prodotto almeno 2 volte e incontrare così i criteri di attestazione validi per la stesura degli Inventari Fonetici proposta da Stoel-Gammon (1985). A questo proposito abbiamo considerato che la procedura relativa alla raccolta di dati fonetici per la compilazione di un inventario è diversa da quella relativa alla raccolta per un test fonetico. Di conseguenza, era sufficiente proporre due parole diverse per lo stesso fono per ciascuna posizione, e il fono poteva ricevere l'attestazione di presenza in base alla produzione anche di una sola parola.
- Poiché restavano pur sempre troppe parole, e considerando che la maggior parte di esse era stata introdotta solo per tener conto della presenza di gruppi consonantici diversi, e poiché questi gruppi consonantici sono moltissimi ma ben difficilmente ciascuno di loro da solo è così frequente da essere presente in almeno due parole, abbiamo deciso di semplificare drasticamente la presenza dei gruppi consonantici, seguendo 2 strategie:

- presentare i gruppi consonantici omosillabici solo in posizione iniziale, e riservare la posizione mediana solo ai gruppi eterosillabici;
- testare la capacità di produrre nessi consonantici, proponendo non tanto la sequenza degli stessi due o tre foni in due parole diverse, ma la stessa Classe Fonologica Naturale, che può essere rappresentata da foni diversi. Per es., in posizione iniziale [**pj**Ede], [**kj**ave] e [gwanto] sono tre nessi consonantici diversi, ma si possono classificare tutti e tre come OCCLUSIVA+ SEMI-CONSONANTE. Allo stesso modo, in posizione mediana, [bi**mb**i], [de**nt**i] e [ga**mb**e] sono tre nessi consonantici diversi, ma si possono classificare tutti come NASALE+ OCCLUSIVA.

Per quanto riguarda *Phon*, vorremmo spezzare una lancia a favore della grande importanza che il suo uso possa assumere nel campo dello studio del linguaggio, per tutti quei settori in cui è essenziale e necessario il confronto tra il target fonetico/fonologico e la produzione effettiva del soggetto, che molte volte, per le ragioni più varie, può non essere corretta. Questo è il caso dello sviluppo fonetico/fonologico normale, in cui i bambini attraversano stadi di sviluppo linguistico che contemplano la semplificazione dei target lessicali, ma è anche il caso dell'acquisizione di L2 da parte di parlanti anche adulti, in cui il discente modifica la produzione del target per l'interferenza delle regole della sua lingua madre, ed è anche il caso del parlato cosiddetto patologico, in cui il target viene modificato a causa della presenza di stati patologici di origine centrale e/o periferica. È facile comprendere allora come questo strumento possa offrire un potente aiuto al logopedista che nella sua pratica clinica quotidiana viene costantemente in contatto con tutte e tre le tipologie descritte. Ricordiamo infatti come tra le funzioni di questo potente software si annovera:

- la connessione ai dati multimediali (per una verifica costante della trascrizione col dato acustico, ma anche per la possibilità di analizzare quest'ultimo con software specifici come PRAAT),
- la segmentazione delle unità in più livelli (dalla parola, alla sillaba, al fonema),
- la trascrizione in doppio cieco (per effettuare una corretta trascrizione e per verificarne l'attendibilità ricavando l'indice di concordanza tra i trascrittori),
- l'etichettatura automatica dei dati (che avviene ad esempio durante la sillabazione dei target e delle produzioni effettive),
- la comparazione sistematica tra il target fonologico (modello) e le forme fonetiche (effettivamente prodotte), reso possibile dall'allineamento automatico.

Inoltre, *Phon* offre i vantaggi tipici di un database elettronico, cioè rende possibile l'archiviazione in forma stabile e organizzata dell'intera mole di dati inseriti in modo da potervi accedere in qualsiasi momento facendo delle ricerche mirate. Concludiamo ricordando che chi scarica PHON e lo usa per creare database si impegna moralmente verso i suoi creatori a inserire i suoi dati nel database *online* del Consorzio CHILDES, perchè è solo attraverso la condivisione e la discussione pubblica che la conoscenza scientifica può avanzare.

6. BIBLIOGRAFIA

AA. VV. (1999), *Handbook of the International Phonetic Association*, Cambridge: Cambridge University Press.

Bardozzetti, M.P. (2008), *Presentazione ed esemplificazione di 'PHON', un programma per la codifica e l'analisi automatica degli aspetti segmentali del parlato*, Tesi di Laurea in Logopedia, Università degli Studi di Padova, AA 2007-2008.

Bertinetto, P.M. & Loporcaro, M. (2005), The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome, *Journal of the International Phonetic Association*, 35, 131-151.

Bonifacio, S. & Zmarich, C. (2007), *Creazione di un test per la valutazione dello sviluppo fonetico in età precoce,* Progetto di ricerca corrente dell'IRCCS Burlo Garofolo di Trieste, responsabile del progetto: dott.ssa Elisabetta Zocconi.

Bortolini, U. (1995), *PFLI Prove per la valutazione fonologica del linguaggio infantile*, Padova: Edit Master Srl.

Boysson-Bardies, B. de, Vihman, M. M., Roug-Hellichius, L., Durand, C., Landberg, I., & Arao, F. (1992), Material evidence of infant selection from the target language: A crosslinguistic phonetic study. In *Phonological development: Models, research, implications* (C. Ferguson, L. Menn & C. Stoel-Gammon, editors), Parkton, MD: York Press.

Caselli, M.C. & Casadio, P. (1995), *Il primo vocabolario del bambino*, Milano: Franco Angeli.

Caselli, M.C., Pasqualetti, P. & Stefanini, S. (2007), *Parole e frasi nel 'Primo vocabolario del bambino'*, Milano: Franco Angeli.

CHILDES: http://childes.psy.cmu.edu/

Curtin, S. & Werker, J.F. (2007), Perceptual Foundations of Phonological Development. In *Oxford Handbook of Psycholinguistics* (M. Gareth Gaskell, G.T.M Altmann, P.Bloom, A. Caramazza & P. Levelt, editors), Oxford University Press.

Desmarais, C., Sylvestre, A., Meyer, F., Bairati, I. & Rouleau, N. (2008), Systematic review of the literature on characteristics of late-talking toddlers, *International Journal of Language and Communication Disorders*, 43, 4, 361-389.

Dinnsen, E. D., A. (1992), Variation in developing and fully developed phonetic inventories, in *Phonological Development. Models, Research, Implications* (C.A.Ferguson, L. Menn & C. Stoel-Gammon, editors), York Press: Timonium, 191-210.

Doimo, L. (1998), Schedatura Prove, in *La valutazione della comunicazione linguistica. Teorie, metodi, prove* (A. Pinton & L. Lena, editors), Imprimenda, 153-166.

Edwards, J. & Beckman, M.E. (2008), Methodological questions in studying consonant acquisition, *Clinical Linguistics & Phonetics*, 22(12), 937-956.

Ingram, D. (1981), *Procedures for the phonological analysis of children's language*, Baltimore: University Park Press.

Kresheck, J. & Socolofsky, G. (1972), Imitative and spontaneous assessment of 4-year-old children, *Journal of Speech and Hearing Research*, 15, 729–733.

MacNeilage, P.F. & Davis, B.L. (2000), On the origin of Internal Structure of Word Forms, *Science*, 288, 527-531.

Muljacic Z. (1972), *Fonologia della lingua italiana*, Bologna: Il Mulino.

Pedrabissi L. & Santinello M. (1997), *I test psicologici*, Bologna: Il Mulino.

Phon: http://phon.ling.mun.ca/phontrac

Rosolen, D., Barca, A. (1999), Applicazione ed uso dello strumento PFLI, in *La specificità logopedica: valutazione e bilancio* (L. Borgo, editor), Tirrenia (Pisa): Edizioni Del Cerro, 218-223.

Shriberg, L.D & Kwiatkowski, J. (1985), Continuous speech sampling for phonologic analyses of speech-delayed children, *Journal of Speech and Hearing Disorders*, 50, 323-334.

Smit, A.B., Hand, L., Freilinger, J.J., Bernthal, J.E. & Byrd, A. (1990), The Iowa articulation norms project and its nebraska replication, *Journal of Speech and Hearing Disorders*, 55, 779-789.

Smit, A.B. (1993a), Phonologic error distributions in the Iowa- Nebraska articulation norms project: consonant singletons, *Journal of speech and hearing research*, 36, 533-547.

Smit, A.B. (1993b), Phonologic error distributions in the Iowa-Nebraska articulation norms project: word-initial consonant clusters, *Journal of speech and hearing research*, 36, 931-947.

Stoel-Gammon, C., (1985), Phonetic inventories, 15-24 months: a longitudinal study, *Journal of Speech and Hearing Research*, 28, 505-512.

Vihman, M.M. & Boysson-Bardies, B., de (1994), The nature and origins of ambient language influence on infant vocal production and early words, *Phonetica*, 51, 159– 169.

Zmarich, C. & Bonifacio, S. (2005), Phonetic inventories in Italian children aged 18-27 months: a longitudinal study, in *Proceedings of INTERSPEECH'2005-EUROSPEECH*, Lisboa, September 4-8, 757-760.

Zmarich, C., Dispaldro, M, Rinaldi, P. & Caselli, M. C. (2009), La composizione fonetica del primo vocabolario del bambino, in *La Fonetica Sperimentale: Metodo e Applicazioni,* Atti del 4° convegno AISV, Università della Calabria, 3-5 dicembre 2007 (L. Romito, V. Galatà & R. Lio, editors), Torriana (RN): EDK Editore, 324-336.

Zmarich, C., Dispaldro, M., Rinaldi, P. & Caselli, M.C. (accettato per la pubblicazione), Caratteristiche fonetiche del 'Primo Vocabolario del Bambino', *Psicologia Clinica dello Sviluppo*.

7. APPENDICE 1: lista delle parole contenute nel test applicato al campione di questo studio.

Item lessicali	
Onomatopee	baubau, cipcip, miao, tuttu.
Nomi	cane, capra, cavallo, coccodrillo, coniglio, elefante, farfalla, gallina, gallo, gatto, giraffa, ippopotamo, leone, lupo, maiale, mosca, orso, papera, pecora, pesce, pinguino, pulcino, rana, scimmia, scoiattolo, tartaruga, tigre, uccellino, zanzara, zebra, auto, barca, bicicletta, camion, elicottero, moto, passeggino, treno, bambola, cubo, dado, lego, palla, palloncino, pistola, secchiello, tromba, trottola, acqua, arancia, banana, biscotti, caramella, ciliegie, cioccolata, cracker, formaggio, fragola, latte, leccalecca, mela, noccioline, pane, pappa, piselli, pizza, pomodoro, riso, succo, torta , uva, bavaglino, calze, collana, guanti, occhiali, pantaloni, scarpe, sciarpa, stivali, bocca, braccio, capelli, denti, dito, faccia, gambe, ginocchio, guance, labbra, lingua, mano, naso, occhio, ombelico, pancia, piede, sederino, unghie, asciugamano, biberon, bicchiere, bottiglia, candelina, chiave, ciuccio, coltello, coperta, cucchiaio, cuscino, dentifricio, fazzoletto, forbici, forchetta, giornale, libro, matita, martello, ombrello, orologio, pentola, pettine, piatto, sacchetto, sapone, scatola, scopa, shampoo, soldi, spazzolino, specchio, tappo, tazza, telefono, vasino, cassetto, scala, sedia, tavolo, albero, campana, fiore, foglia, luna, sasso, scivolo, sole, stella, casa, giostra.
Persone	bimbi, mamma, nonna, nonno, papà, tati, vigile.
Routine	ciao, nanna, no.
Aggettivi	rosso, verde.

8. APPENDICE 2: esempio di lista di parole distribuite in funzione del fono che da elicitare per i bambini della prima fascia (18-23 mesi) e applicata al campione di 12 bambini analizzati in questo studio.

Posizione	p	b	t	d	k	g	m	n	f	v	s	z	S
iniziale	**p**alla **p**ane **p**appa **p**a'pa **p**omodoro	**b**aubau **b**anana **b**iskotti **b**iberon **b**imbi **b**okk**a**	**t**u'tu **t**ati	**d**enti **d**ito **d**ado	**k**ane **k**apelli **k**ubo	**g**atto **g**ambe	**m**ela **m**ano **m**atita **m**amma	**n**azo **n**onna **n**onno **n**anna **n**o			**s**ukko		
iniziale gr. cons.	**pj**ede				**sk**arpe	**gw**anti	**mj**ao		**fr**agola		**sk**arpe		
mediana	pa**pp**a ka**p**elli pa'**p**a	bi**b**eron ku**b**o bau**b**au	tu'**t**u biskо**tt**i la**tt**e di**t**o mati**t**a ta**t**i ga**tt**o	pje**d**e pomo**d**oro da**d**o	su**kk**o bo**kk**a	tartaru**g**a le**g**o fra**g**ola	ma**mm**a po**m**odoro	ba**n**a**n**a pa**n**e ma**n**o no**nn**a no**nn**o na**nn**a				na**z**o	
mediana gr. cons.	skar**p**e	bi**mb**i ba**wb**au ga**mb**e	de**nt**i gwa**nt**i		a**kw**a bi**sk**otti o**kkj**o		bi**mb**i ga**mb**e	de**nt**i gwa**nt**i			bi**sk**otti		

Posizione	ts	dz	tS	dZ	l	L	r	j	w
iniziale			**tS**utS:o **tS**ao		**l**ego **l**atte				
iniziale gr. cons.									**gw**anti
mediana			tSu**tt**S**o**	orolo**dZ**o	pa**ll**a me**l**a kape**ll**i orolodZo fragо**l**a		bibe**r**on o**r**olodZo pomodo**r**o		
mediana gr. cons.							skar**p**e	o**kkj**o	a**kw**a

ENFASI E CONFINI PROSODICI IN DUE STILI DI ELOQUIO EMOZIONALE

Pier Luigi Salza, Enrico Zovato, Morena Danieli
Loquendo S.p.A. – Torino
pierluigi.salza@loquendo.com, enrico.zovato@loquendo.com, morena.danieli@loquendo.com

1. SOMMARIO

La nuova frontiera delle tecnologie di sintesi del parlato è la capacità di generare segnali vocali con caratteristiche espressive in grado di variare in modo analogo a quanto avviene nella voce umana. Molti studi hanno evidenziato correlazioni significative tra lo stile di eloquio e le variazioni di alcuni parametri prosodici e spettrali; inoltre, alcuni sistemi di sintesi vocale sono in grado di replicare in parte queste variazioni (Schröder, 2001). Nel presente lavoro si illustra uno studio pilota condotto su alcuni fenomeni di natura ritmica e intonativa, relativamente a due stili espressivi, a partire da registrazioni effettuate in laboratorio da parlanti madrelingua inglesi. Si sono annotati, in ciascun enunciato, i confini prosodici e i fenomeni di enfasi eventualmente realizzati (le parole, o i gruppi di parole, prominenti). L'obiettivo di questo studio pilota è cercare di individuare eventuali correlazioni tra i fenomeni prosodici presi in considerazione e lo stile di eloquio. Lo scopo applicativo è il tentativo di riprodurre, nel sistema di sintesi vocale, analoghi meccanismi per una più accurata caratterizzazione in senso espressivo dei segnali generati. In particolare, si potrebbero introdurre, nel modulo di assegnazione automatica del *phrasing*, regole dipendenti dallo stile e, nel contempo, generare segnali con realizzazioni acustiche dei fenomeni di prominenza, anch'essi legati in modo contestuale allo stile adottato. Gli stili emozionali studiati in questo progetto sono quelli corrispondenti allo stile triste e allo stile allegro. La scelta di ridurre lo studio a due soli stili è stata dettata da motivazioni pratiche legate anche ai domini di applicazione dei prototipi che sono oggetto di sviluppo nell'ambito del progetto citato. Queste applicazioni mirano alla realizzazione di un assistente virtuale capace di conversare con l'utente in modo affettivo, riproducendo, tramite la voce e altre modalità, comportamenti emozionali legati al contesto del dialogo. L'agente deve pertanto essere capace di assumere un atteggiamento positivo o negativo a seconda delle reazioni dell'utente.

2. INTRODUZIONE

La ricerca nel campo della sintesi del parlato ha ormai più di trent'anni: i sistemi di sintesi attuali hanno raggiunto traguardi impressionanti in termini di naturalezza e di intelligibilità del segnale vocale, senza tuttavia che l'obiettivo di poter generare segnali sintetici indistinguibili dalla voce umana in termini di variabilità espressiva sia ancora stato realizzato. Ed è proprio su questo campo, vale a dire sulla capacità di un sistema di sintesi di generare segnali che possano variare quanto all'espressività e al 'colore emozionale', che si apre oggi una nuova sfida per gli sperimentatori nel campo della sintesi da testo, una sfida che ha portato in primo piano l'indagine sui parametri prosodici (ritmo, intonazione, velocità di eloquio e enfasi) e su quelli spettrali del parlato, che da lungo tempo gli studiosi di fonetica hanno individuato come terreno in cui è possibile rintracciare correlazioni significative con le variazioni dello stile di eloquio dei parlanti (Scherer, 2003; Johnstone &

Scherer, 1999). Particolarmente numerose sono le ricerche in campo fonetico che sottolineano la complessità dello studio del parametro più rilevante, l'F0 (si veda, tra gli altri, Magno Caldognetto, 2002).

Al fiorire degli studi in questo campo hanno contribuito anche le applicazioni dei sistemi di sintesi all'interno di interfacce persona-macchina sempre più sofisticate: numerosi studi hanno infatti dimostrato che un'intonazione appropriata migliora in modo significativo la qualità percepita della conversazione con un agente conversazionale (Poggi *et al.*, 2005; Bevacqua *et al.*, 2007) e che la voce svolge un ruolo cruciale nel rivelare lo stato emotivo del parlante e nel comunicare empatia con lo stato emozionale del partner conversazionale (Campbell, 2005).

Il presente studio pilota si inserisce in questo campo di ricerca e ha l'obiettivo di analizzare la variazione di alcuni fenomeni di natura intonativa e ritmica relativamente a due stili espressivi, a partire da registrazioni effettuate in laboratorio da parte di parlanti madrelingua inglesi. L'analisi acustica è finalizzata all'individuazione di eventuali correlazioni significative tra l'occorrenza dei fenomeni prosodici presi in considerazione e lo stile di eloquio del parlante. Lo scopo applicativo è tentare di produrre, nel sistema di sintesi vocale, una più accurata caratterizzazione in senso espressivo dei segnali generati, attraverso l'introduzione, nel modulo di assegnazione automatica del *phrasing*, di regole dipendenti dallo stile e, nel contempo, la generazione di segnali con realizzazioni acustiche dei fenomeni di prominenza, anch'essi legati in modo contestuale allo stile adottato.

Gli stili espressivi studiati in questo progetto sono quelli corrispondenti agli stati emozionali triste e allegro. La scelta di ridurre lo studio a due soli stili è stata dettata da motivazioni pratiche legate anche ai domini di applicazione dei prototipi che sono oggetto di sviluppo nell'ambito del progetto più ampio in cui questo lavoro si colloca. Queste applicazioni mirano alla realizzazione di un assistente virtuale capace di conversare con l'utente in modo affettivo, riproducendo, tramite la voce e altre modalità, comportamenti emozionali legati al contesto del dialogo. L'agente dovrebbe pertanto essere capace di dialogare in modo empatico con l'interlocutore, assumendo un atteggiamento allegro o triste a seconda dello stato emotivo dell'utente.

Com'è noto, sono particolarmente diffusi due modelli di rappresentazione delle emozioni: il modello categoriale (Eckman, 1992) e quello dimensionale (Russell, 1980; Davidson *et al.*, 2009). Il primo tratta le emozioni come categorie discrete e riconosce un insieme ristretto di emozioni di base, il secondo fa invece riferimento alla possibilità di far corrispondere ad una data emozione un grado di valenza e uno stato di attivazione della stessa in qualche modo 'misurabili'. Mentre in psicologia i due modelli fanno riferimento ad assunti teorici, strutture concettuali e metodologie sperimentali piuttosto diverse, nel campo della letteratura sulla sintesi da testo il riferimento a un modello o all'altro non implica necessariamente, tranne in pochi casi, per esempio (Scherer, 2003), che gli autori si schierino con una particolare teoria sulla natura delle emozioni, né che rinuncino ad utilizzare con una certa libertà i diversi paradigmi sperimentali. In questo studio pilota si utilizzano etichette categoriali ('triste' e 'allegro') per riferirsi a stili di eloquio le cui realizzazioni possono essere descritte in relazione all'andamento di parametri puramente acustici.

3. SCELTA E VALUTAZIONE DEI MATERIALI

L'esperimento è focalizzato su due stili espressivi indotti da stati emotivi con valenza opposta: positiva vs. negativa e attivazione medio-alta nel primo caso e bassa nel secondo. In termini di rappresentazione categoriale si possono etichettare i due stili come *allegro* e *triste*.

I dati per l'esperimento sono stati raccolti mediante registrazioni effettuate appositamente in laboratorio da tre parlanti madrelingua inglesi: un parlante maschile madrelingua britannico (M1), e due parlanti madrelingua americani, uno maschile (M2) e uno femminile (F1). Al fine di ottenere del materiale vocale con due stili espressivi ben caratterizzati, un esperto madrelingua inglese ha selezionato, da romanzi e racconti, testi il cui contenuto semantico fosse in grado di indurre nel lettore il corretto atteggiamento emotivo. Si sono ottenuti due insiemi di testi, uno per ciascuno stile considerato, costituiti da frasi più o meno lunghe, fino all'estensione del breve paragrafo. Ecco due esempi di testi nei due stili considerati:

- stile allegro: "For instance, on the planet Earth, man had always assumed that he was more intelligent than dolphins, because he had achieved so much, the wheel, New York, wars and so on" (da *The Hitch Hiker's Guide to the Galaxy* di Douglas Adams);
- stile triste: "Life's but a walking shadow, a poor player that struts and frets his hour upon the stage, and then is heard no more" (da *Hamlet* di William Shakespeare).

Durante le registrazioni si è chiesto ai tre parlanti di leggere i brani in modo appropriato e naturale, conformemente ai contenuti. Si è poi chiesto ai parlanti di leggere gli stessi brani, in sessioni di registrazione separate, adottando uno stile di lettura neutro indipendentemente dai contenuti semantici dei testi, ottenendo così un corpus di riferimento con stile neutro.

Tutte le registrazioni sono state effettuate in camera silente con dispositivi di acquisizione di alta qualità. I file audio sono stati convertiti in forma numerica con frequenza di campionamento di 44,1 Khz, a 16 bit e con codifica PCM lineare. Per comodità di elaborazione, tutti i testi sono stati suddivisi in frasi di lunghezza variabile, tutte terminanti con '.' (conclusive) e non eccedenti le 50 parole.

Per ciascun parlante, il corpus originale consiste in 42 frasi espressive (14 in stile allegro e 28 in stile triste) e in 42 corrispettive versioni neutre. In totale, dunque, il corpus ammonta a (42+42) x 3 = 252 frasi. Il corpus è stato sottoposto ad una valutazione percettiva tesa a verificare l'effettiva resa emozionale. A questo fine è stata approntata un'interfaccia WEB tramite la quale lo stile emozionale di ciascuna frase di ciascuno speaker (incluse le versioni neutre) è stato valutato percettivamente da parte di 3 soggetti madrelingua inglesi.

Le frasi sono state ascoltate in ordine casuale per quanto riguarda lo stile espressivo. Ciascun enunciato poteva essere riascoltato più volte. Infine era richiesta una valutazione su una scala MOS (*Mean Opinion Score*) a 7 punti, avente ad un estremo 1: very sad e all'altro estremo 7: very happy. La finestra dell'interfaccia è mostrata nella figura 1.

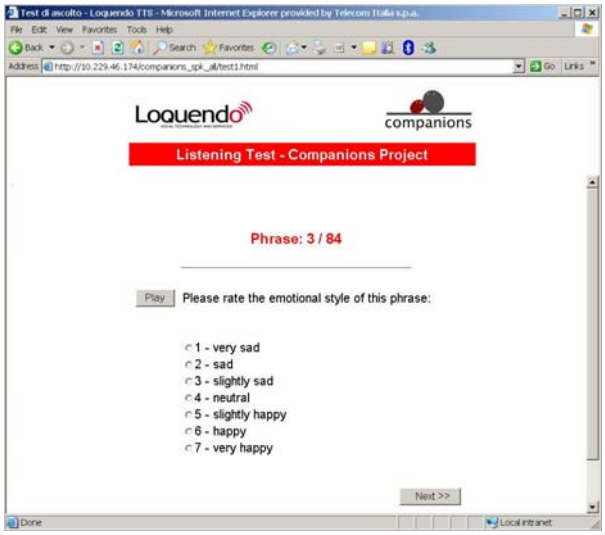


Figura 1: Finestra dell'interfaccia WEB per la valutazione soggettiva dello stile emozionale delle frasi registrate

Per discriminare i dati espressivi da quelli neutri, si sono calcolate le medie dei punteggi ottenuti da ciascuna frase. Come si può osservare dai risultati mostrati in figura 2, una percentuale significativa delle valutazioni si concentra sui punteggi 3 e 5 (*slightly sad* e *slightly happy*, rispettivamente), ma una parte non trascurabile si concentra sul punteggio centrale 4 (*neutral*).

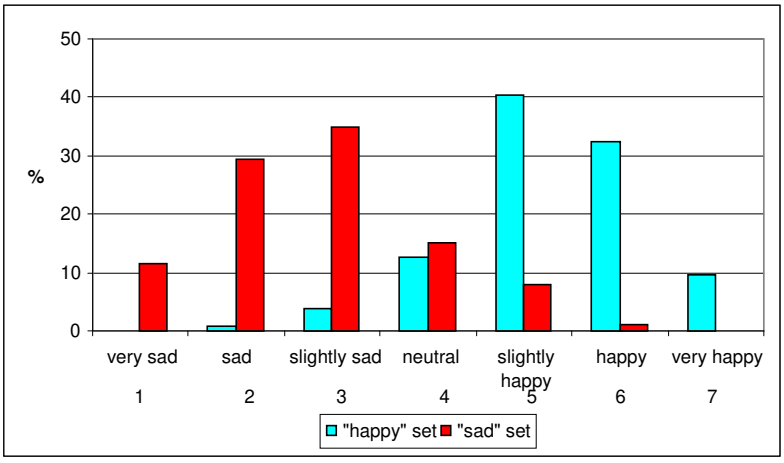


Figura 2: Distribuzione dei punteggi medi, calcolati su 3 giudizi, per frasi tristi e allegre e neutre

Le frasi che hanno ottenuto un punteggio medio, calcolato su 3 giudizi, tra 3,5 e 4,5, sulla base della scala di valutazione utilizzata, sono state incluse nello stile neutro; quelle con un punteggio medio superiore a 5 sono state assegnate allo stile allegro; quelle con un punteggio medio inferiore a 3 allo stile triste.

Si è così ricavato complessivamente un sottoinsieme di 176 frasi, di cui 90 effettivamente espressive e 86 decisamente neutre.

4. ANNOTAZIONE E SELEZIONE DEI DATI DA ANALIZZARE

I testi corrispondenti agli enunciati selezionati sono stati annotati manualmente e separatamente da due operatori che, per ciascuna frase appartenente ai tre insiemi (allegro, triste e neutro), hanno effettuato l'ascolto (ripetibile a volontà anche su singole porzioni dell'enunciato) ed hanno poi etichettato ciascun confine di parola percepito come confine prosodico[1] e ciascuna parola percepita come avente la presenza di enfasi[2]. L'annotazione dell'enfasi è stata condotta su base percettiva. Gli operatori sono stati addestrati a segnalare ciascuna parola sulla quale fosse percepibile una discontinuità rispetto al contesto precedente e/o seguente (rallentamento o accelerazione della velocità di eloquio e/o variazione dell'intonazione e/o particolare inflessione intontiva e/o variazione dell'intensità) tale da far ritenere la parola come pronunciata in maniera prominente, marcata o enfatica.

Nel seguito si riportano tre esempi di frasi annotate (appartenenti rispettivamente agli insiemi allegro, triste e neutro), dove: <P> è l'etichetta usata per marcare un confine prosodico e <EM> quella usata per marcare l' enfasi.

- Stile allegro:
  For <EM instance>,<P> on the planet <EM Earth>,<P> man had <EM always> assumed <P> that he was more <EM intelligent> than dolphins,<P> because he had achieved so <EM much> <P>, the wheel,< P> New York, <P> wars and <EM so on>.
- Stile triste:
  Life's but a walking shadow,<P> a <EM poor> player <P> that <EM struts> <P> and frets his hour upon the stage,<P> and then is <-EM heard> no more.
- Stile neutro:
  Life's but a walking shadow,<P> a poor player that struts and frets his hour upon the stage,<P> and then is heard no more.

La congruenza tra le annotazioni dei due operatori è stata valutata, separatamente per ciascun enunciato e per ciascun parametro (confine prosodico e enfasi), tramite il 'K test'. Il coefficiente *K* viene usato per misurare il livello di concordanza tra le classificazioni di due valutatori secondo determinate categorie, rispetto alla concordanza *per caso*, in base alla seguente espressione (Carletta, 1996):

[1] Con il termine 'confini prosodici' si intendono le cesure nel *continuum* del parlato utilizzate per suddividere l'enunciato in unità tonali (o sintagmi intonativi); non si sono distinti i confini forti (caratterizzati dalla presenza di una pausa) da quelli deboli (caratterizzati solo da movimenti di F0).

[2] Con il termine 'enfasi' si intendono quei fenomeni di prominenza a livello intonativo, di durata e di intensità, tesi ad evidenziare alcune parole nel contesto della frase (focus).

$$K = \frac{P(a) - P(e)}{1 - P(e)}$$

(1)

dove: P(a) è la percentuale di concordanza osservata tra i due valutatori e P(e) è la probabilità di concordanza *per caso* tra i due valutatori.

I valori di K sono compresi tra -1 e 1, dove -1 sta per 'massimo disaccordo', 1 sta per 'perfetta concordanza', 0 sta per "concordanza uguale a quella attesa in base al caso". Per ciascuna frase, il 'K test' è stato calcolato considerando due possibili stati di ciascun confine di parola della frase, vale a dire presenza o assenza di confine prosodico, e due possibili stati di ciascuna parola della frase, dati dalla presenza o assenza di enfasi. Ottenuto il valore di K per ciascuna frase annotata, e per ciascun parametro, si è ulteriormente selezionato il materiale su cui svolgere le successive analisi sulla base del livello di concordanza. Si sono prese in considerazione soltanto quelle frasi il cui valore di 'K' è superiore a 0,6, che, secondo (Landis & Koch, 1977), significa 'concordanza sostanziale' (si veda la Tabella 1). Dopo questa ulteriore selezione si è ottenuto il corpus definitivo, composto da due corpora parzialmente sovrapposti, uno per i confini di parola, l'altro per le enfasi. Essi contano rispettivamente: <P>: 76 enunciati espressivi (35 allegri e 41 tristi) e 66 enunciati neutri; <EM>: 73 enunciati espressivi (34 allegri e 39 tristi) e 48 enunciati neutri. A causa di questo complesso processo di selezione, il set finale dei dati non risulta omogeneamente distribuito tra i diversi parlanti, quanto a numero e tipologia espressiva delle frasi. Tuttavia, per l'obbiettivo di questo lavoro, ciò non rappresenta una criticità.

K	***LIVELLO DI CONCORDANZA***
< 0	Nessuna
0,0 - 0,20	Leggera
0,21- 0,40	Sufficiente
0,41- 0,60	Moderata
0,61- 0,80	Sostanziale
0,81-1,00	Perfetta

Tabella 1: Valore di K e livello di concordanza (cfr. Landis & Koch, 1977)

I valori medi di K per <P> ed <EM>, separatamente per i tre parlanti, sono riportati nella Tabella 2. Da questa analisi si evince che la concordanza tra le due annotazioni è stata buona, con valori medi superiori a 0,6.

Kappa	<P> Allegro	<P> Triste	<P> Neutro	<EM> Allegro	<EM> Triste	<EM> Neutro
F1	0,71	0,84	0,73	0,84	0,68	0,38
M1	0,88	0,69	0,70	0,80	0,97	0,64
M2	0,85	0,76	0,78	0,74	0,67	0,60

Tabella 2: Valori medi di K per gli enunciati espressivi e neutri dei tre parlatori

I dati utilizzati per la successiva analisi sono quelli relativi alle frasi in cui si è verificata sostanziale concordanza tra le due annotazioni, ovvero i casi in cui k>0,6. Come evidenziato dai grafici riportati nella figura 3, i casi in cui si riscontrano valori di Kappa maggiori di questa soglia sono la maggioranza. Infatti, ad eccezione delle annotazioni delle enfasi delle frasi in stile neutro della speaker F1, tutti gli altri casi mostrano percentuali di superamento di questa soglia ben oltre il 50%.

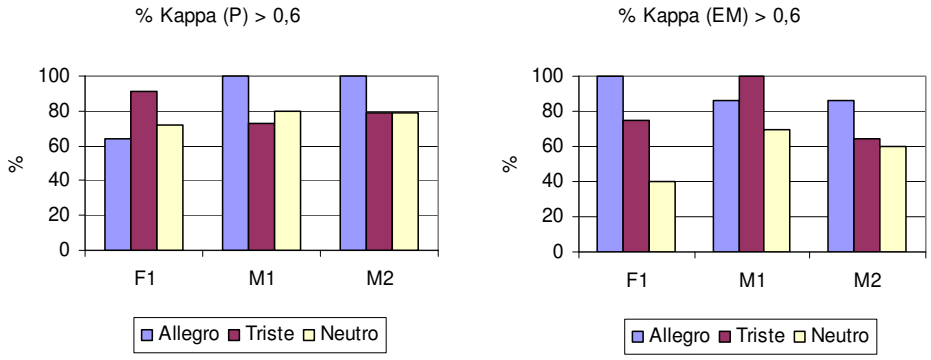


Figura 3: Percentuali di casi in cui il valore di K delle annotazioni è maggiore della soglia 0,6

5. RISULTATI

Il numero di eventi è stato rapportato alla lunghezza della frase. Per ciascuna frase sono stati calcolati due indici della frequenza di occorrenza dei fenomeni in esame, C_P e C_{EM}, dati dal rapporto, rispettivamente, tra il numero dei confini prosodici annotati e il numero di confini di parola e tra il numero di enfasi annotate e il numero di parole, vale a dire:

(2) $$C_P = \frac{N < P >}{NConfini\_Parola}$$

(3) $$C_{EM} = \frac{N < EM >}{NParole}$$

Analizzando le tre serie di valori di ciascun indice C_P e C_{EM}, ognuna corrispondente ad uno dei tre stili considerati (allegro, triste e neutro), si osserva che i valori medi sono differenti nei tre casi. I grafici mostrati nelle figure 4 e 5 riportano i valori medi dei coefficienti C_P e C_{EM}, disaggregati per valutatore, per parlante e per stile di eloquio. Per quanto concerne i confini prosodici si nota che solo uno dei parlanti tende ad aumentare il numero di confini nello stile triste, come evidenziato in figura 4. Sarà necessario in futuro approfondire l'analisi distinguendo i confini forti da quelli deboli (vale a dire con o senza pausa realizzata). Riguardo all'enfasi risulta invece una maggiore coerenza tra i tre parlanti. Il fenomeno appare più frequente nello stile allegro e meno in quello triste, in cui comunque è significativamente più presente rispetto allo stile neutro, come mostrato in figura 5.

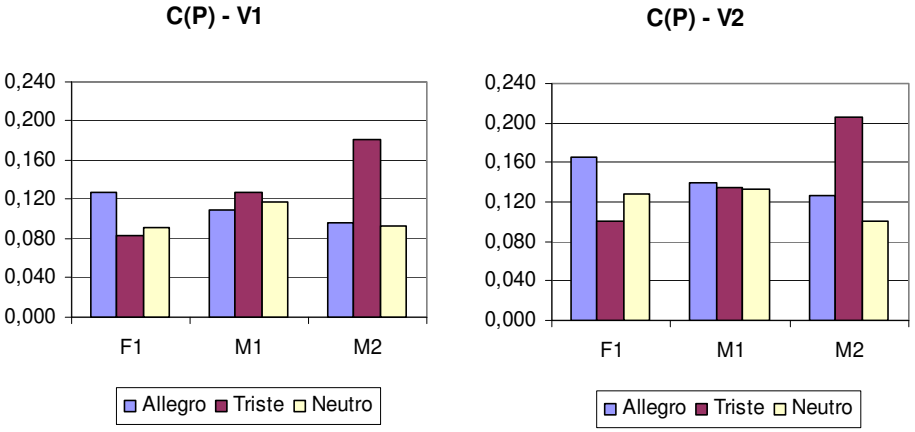


Figura 4: Valore medio del coefficiente C_P, calcolato separatamente sui dati dei due valutatori

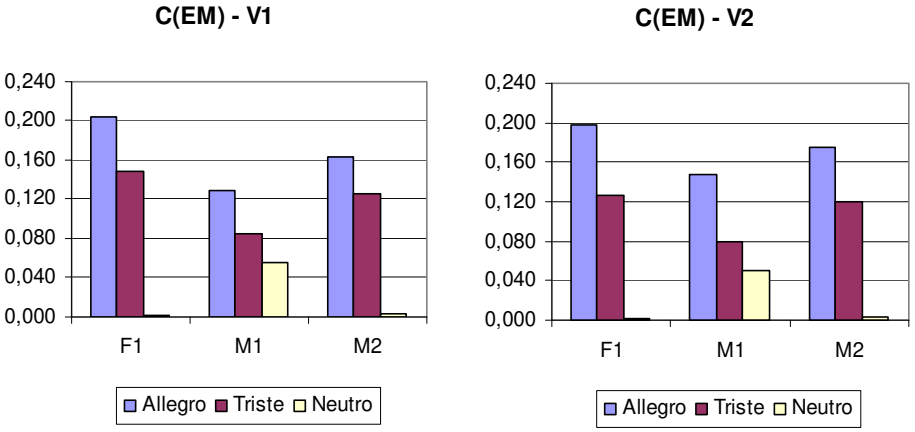


Figura 5. Valore medio del coefficiente C_{EM}, calcolato separatamente sui dati dei due valutatori.

La sostanziale concordanza tra le annotazioni dei due valutatori si riflette anche nei valori medi dei coefficienti di frequenza (figure 4 e 5). In considerazione di ciò, la successiva analisi statistica e della morfologia si è basata sulle annotazioni di uno solo dei due valutatori.

5.1 Analisi statistica dei dati

Al fine di valutare la significatività dei dati riportati nel precedente paragrafo, è stata effettuata un'analisi statistica ANOVA su serie di dati relativi ai coefficienti C_P e C_{EM} classificate per stile di eloquio. Tutti i dati relativi ad uno dei due valutatori e la cui concordanza con l'altro valutatore ha fornito valori di K>0,6, sono stati raggruppati per stile e ordinati per valori di k decrescenti. Per l'analisi statistica sono state usate le frasi corrispondenti ai primi 34 valori di queste serie dei coefficienti C_P e C_{EM}. Il numero di coefficienti è stato scelto in modo da garantire la presenza di una quantità adeguata di dati di ciascun speaker. Il risultato dell'analisi conferma una variazione significativa della frequenza delle enfasi negli stati allegro e triste rispetto allo stile neutro.[3]

Mentre per quanto riguarda i confini prosodici i dati confermano la maggiore incertezza, già evidenziata dalla precedente analisi. Lo stile triste sembra differenziarsi in modo più significativo rispetto allo stile neutro e allegro, con valori più elevati del coefficiente C_P e quindi con una frequenza maggiore di pause.[4]

Nelle figure 6 e 7 sono riportati i *box-plot* riassuntivi per le tre serie di dati relativi allo stile allegro, triste e neutro e per ciascun coefficiente analizzato. Anche dai grafici emergono le differenze nei tre stili, per quanto riguarda il fenomeno di enfasi. Per ciascuna serie sono visualizzati, tra i vari parametri, il valore mediano, il primo e il terzo quartile. Le deviazioni sono indicate con il carattere '+'.

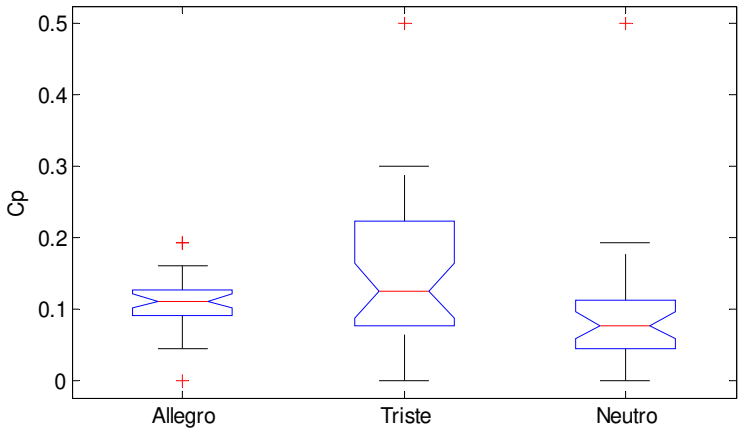


Figura 6: *Box-plot* relativo alle serie dei coefficienti C_P negli stili allegro, triste e neutro

[3] N_{Triste}=34, $N_{Allegro}$=34, N_{Neutro}=34; $F_{Triste\text{-}Neutro}$=58,57 e p<0,00001; $F_{Allegro\text{-}Neutro}$=219,56 e p< 0,00001.

[4] N_{Triste}=34, $N_{Allegro}$=34, N_{Neutro}=34; $F_{Triste\text{-}Neutro}$=4,91 e p<0,03; $F_{Allegro\text{-}Neutro}$=8,17 e p< 0,005.

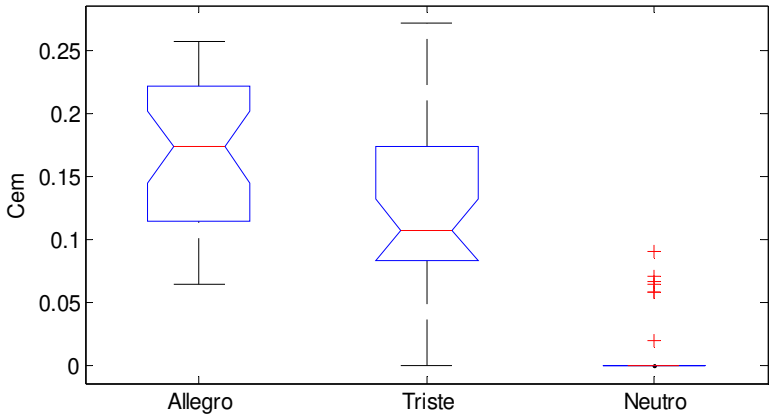


Figura 7: *Box-plot* relativo alle serie dei coefficienti C_{EM} negli stili allegro, triste e neutro

5.2 Morfologia dell'enfasi

Sono state analizzate alcune realizzazioni acustiche dell'enfasi, in particolare per quanto riguarda i valori di F0 ed è stato notato che spesso il dominio della prominenza è inferiore alla dimensione di parola tendendo a localizzarsi sulla sillaba tonica. Si è proceduto ad esaminare i nuclei vocalici delle sillabe toniche nelle parole oggetto di enfasi. In quasi tutti i casi si verifica un picco intontivo e nello stile allegro questo movimento risulta più pronunciato con valori massimi di F0 più accentuati. Nello stile triste, come è lecito attendersi, i massimi di F0 sono di ampiezza minore, coerentemente con i valori mediamente più bassi di F0 che caratterizzano questo stile di eloquio (Scherer, 2003). Per ottenere una misura di queste variazioni è stata preliminarmente condotta un'analisi segmentale sui dati a disposizione. Sono stati individuati nelle forme d'onda i confini fonetici mediante tecniche di allineamento forzato delle etichette fonetiche derivanti dalle trascrizioni dei testi letti (Brugnara *et al.*, 1993). I segnali sono quindi stati elaborati con strumenti automatici che forniscono anche il calcolo dei parametri acustici e in particolare della frequenza fondamentale F0.

La nostra analisi si è quindi concentrata sulle vocali toniche sia in parole oggetto di enfasi che in parole non enfatizzate. In figura 8 sono riportate le variazioni medie dei valori massimi di F0 delle vocali toniche oggetto di enfasi rispetto a vocali toniche in cui non sono state annotate enfasi. Sia nello stile allegro che in quello triste si registrano variazioni positive di questo parametro anche se, come anticipato, di entità decisamente superiore nello stile allegro. E' da notare, inoltre, come queste caratteristiche, seppur di entità diverse, siano comuni a tutti e tre i parlanti.

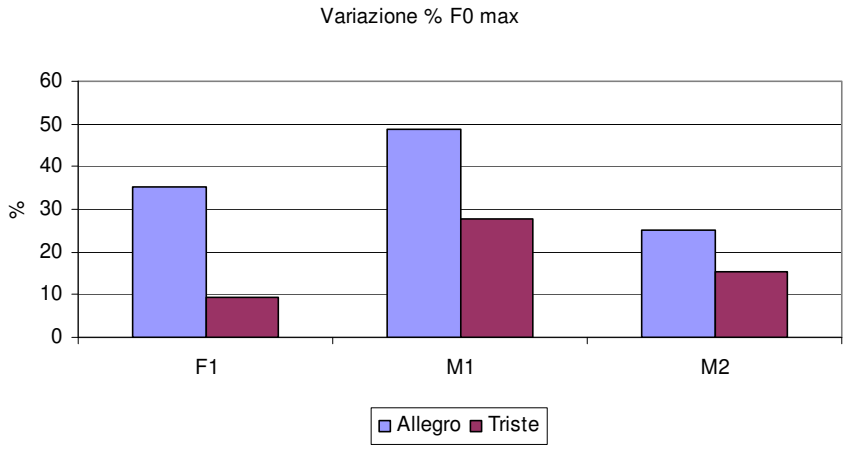


Figura 8: Variazioni percentuali dei valori massimi di F0 nelle vocali toniche di parole enfatizzate rispetto alle vocali toniche di parole non enfatizzate

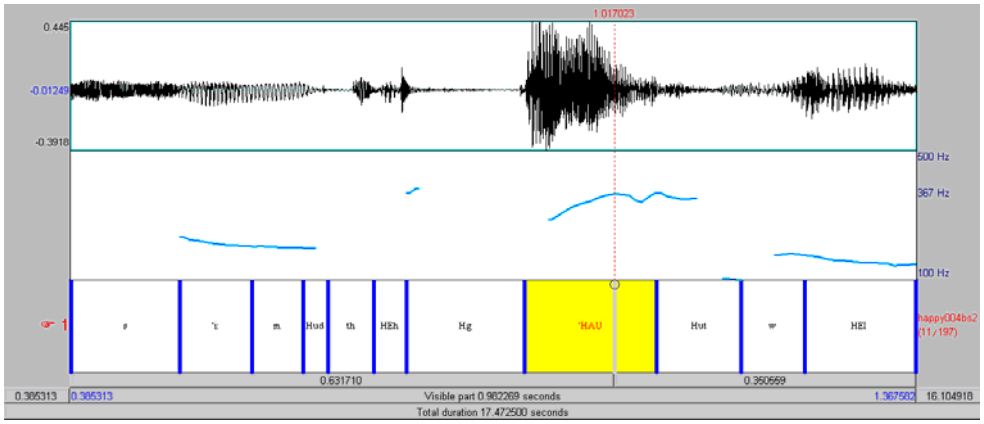


Figura 9: Esempio di forma d'onda, curva di F0 e segmentazione fonetica della sequenza: '...seemed to **out**weigh...' (parlante femminile)

6. CONCLUSIONI E SVILUPPI FUTURI

In questo lavoro si sono presentati i primi risultati di uno studio pilota finalizzato all'analisi delle eventuali correlazioni esistenti tra alcuni fenomeni di tipo prosodico e lo stile di eloquio. In particolare sono state analizzate le occorrenze, percepite da un ascoltatore all'interno di un enunciato, dei seguenti eventi prosodici: enfasi, intesa come fenomeno di prominenza principalmente intontiva (ma anche ritmica e di intensità); confini prosodici, caratterizzati o no da una pausa. Sono stati considerati due stili di eloquio espressivo, che in termini categoriali si possono definire come allegro e triste. Oltre a questi due stili è stato considerato uno stile di eloquio neutro per poter eseguire i necessari confronti. I dati acustici acquisiti da tre parlanti inglesi sono stati annotati da due valutatori,

su base percettiva. Per quanto riguarda l'enfasi, dall'analisi effettuata su una selezione di tali annotazioni si sono potute ricavare alcune osservazioni, che appaiono interessanti pur nella limitatezza dei dati disponibili. La frequenza di occorrenza dell'enfasi sembrerebbe essere superiore nello stile allegro rispetto allo stile triste. Inoltre, si sono misurate variazioni positive dei valori massimi di F0 delle vocali toniche delle parole oggetto di enfasi rispetto a vocali toniche di parole in cui non sono state annotate enfasi, sia nello stile allegro che in quello triste, benché di entità decisamente superiore nello stile allegro. Per quanto riguarda invece i confini prosodici lo studio non ha fornito chiare correlazioni tra la loro frequenza e lo stile di eloquio. Alla luce di questi primi e parziali risultati, si renderebbero pertanto necessarie ulteriori indagini, su un campione di dati acustici più esteso. Per quanto riguarda i confini prosodici, potrebbe apparire utile distinguere tra confini forti e confini dovuti a movimenti intonativi che non implichino l'interruzione del parlato. Riguardo all'enfasi, potrà essere sicuramente interessante un'analisi di tutti i parametri acustici (oltre ad F0, anche durata e intensità) delle sillabe toniche, confrontando quelle delle parole su cui è stata percepita l'enfasi con quelle delle altre parole. Infine, sarà presa in considerazione anche l'eventuale correlazione tra presenza di enfasi e/o confine prosodico e la categoria grammaticale delle parole.

RINGRAZIAMENTI

La ricerca è stata parzialmente finanziata dalla Commissione Europea nell'ambito del progetto FP6 IST34434 'Companions'. Gli autori sono riconoscenti a Kim Bao Nguyen, Simon Parr e Cristina Segatto.

7. BIBLIOGRAFIA

Bevacqua, E., Mancini, M., Niewiadomski, R., & Pelachaud, C. (2007), An expressive eca showing complex emotions, in *Proceedings of the Artificial and Ambient Intelligence convention 2007 (AISB'07)*, Newcastle upon Tyne: Newcastle University.

Brugnara, F., Falavigna, D., & Omologo, M. (1993), Automatic Segmentation and Labelling of Speech based on Hidden Markov Models, *Speech Communication,* 12, 357-370.

Campbell, N. (2005), Getting to the heart of the matter: Speech as the expression of affect; rather than just text or language, *Language Resources and Evaluation*, 39, 109-118.

Carletta, J. (1996), Assessing agreement on classification tasks: the kappa statistics, *Computational Linguistics,* 22, 249-254.

Davidson, R.J, Scherer, K.R., & Hill Goldsmith, H. (Editors) (2009), *Handbook of Affective Sciences*, Oxford (UK): Oxford University Press.

Ekman, P. (1993), An argument for basic emotions, *Cognition and Emotion*, 6, 169-200.

Johnstone, T. & Scherer, K.R. (1999), The effects of emotions on voice quality, in *Proceedings of the 14th International Congress on Phonetic Science*, San Francisco, 2029-2032.

Landis, J.R. & Koch, G.G. (1977), The measurement of observer agreement for categorical data, *Biometrics,* 33, 159-174.

Magno Caldognetto, E. (2002), I correlati fonetici delle emozioni, *Passioni, emozioni, affetti*. (C. Bazzanella & P. Kobau, editors), McGraw-Hill, 197-213.

Poggi, I. & Magno Caldognetto, E. (2004), Il parlato emotivo. Aspetti cognitivi, linguistici e fonetici, in A*tti del Convegno Italiano parlato* (F. Albano Leoni, F. Cutugno, M. Pettorino & R. Savy, editors), Napoli: D'Auria Editore, CD-Rom.

Russell, J.A. (1980), A circumflex model of affect, in *Journal of personality and social psychology*, 1161-1178.

Scherer, K.R. (2003), Vocal communication of emotion: A review of research paradigms, *Speech Communication*, 40, 227-25.

Schröder, M. (2001), Emotional Speech Synthesis: A Review, in *Proceedings of EUROSPEECH 2001*, Aalborg, 561-564.

UN CORPUS SPERIMENTALE PER LO STUDIO CROSS-LINGUISTICO EUROPEO DELLE EMOZIONI VOCALI

Vincenzo Galatà, Luciano Romito
Laboratorio di Fonetica, Università della Calabria
vgalata@libero.it, luciano.romito@unical.it

1. SOMMARIO

La presente proposta costituisce il primo stadio di una ricerca attualmente in corso, volta allo studio cross-linguistico europeo delle emozioni vocali in quattro lingue: italiano, francese, inglese e tedesco.

Se da un lato si rilevano innumerevoli studi sul parlato emotivo nelle singole lingue, dall'altro gli studi di tipo cross-linguistico e cross-culturale risultano essere assai sparuti (per gli studi cross-linguistici volti all'*encoding* si vedano, ad esempio, Anolli *et al.,* 2008a, 2008b; Braun & Oba, 2007; Kori & Magno Caldognetto, 2003; Piôt, 1999; per quelli volti al *decoding* si vedano, invece, gli studi citati e riportati in Pavlenko, 2005, a cui vanno aggiunti Pell *et al.* 2009; Sawamura *et al.*, 2007; Shochi *et al.*, 2007; Droomey *et al.*, 2005; Tickle, 1999, 2000; Magno Caldognetto & Kori, 1983). Ancora meno sono quelli che hanno affrontato entrambi gli aspetti di *encoding* e *decoding* (ad es. Chung, 1999, 2000; Abelin & Allwood, 2000, 2002; Breitenstein *et al.,* 2001; Thompson & Balkwill, 2006). I motivi sono prevalentemente imputabili alla difficoltà che lo studio delle emozioni vocali impone, difficoltà ulteriormente esacerbate nello studio di tipo cross-linguistico-culturale.

Esaminando i risultati degli studi appena menzionati, si evince come le emozioni vocali siano, alla pari di quelle facciali, riconosciute cross-linguisticamente e cross-culturalmente, con risultati nettamente al di sopra della semplice casualità. Da una meta-analisi condotta da Laukka (2004) volta alla verifica del riconoscimento cross-culturale delle emozioni e alla verifica dell'esistenza di specifici *patterns* acustici della voce per categorie discrete di emozioni, è emerso che: a) l'accuratezza di riconoscimento è superiore a quella della casualità per categorie di emozioni più ampie; b) il *decoding* cross-culturale è inferiore al *decoding* intra-culturale del 7%; c) esistono specifici *patterns* acustici nella voce delle emozioni che vengono utilizzati per comunicare emozioni discrete.

Una ricognizione su ben 64 database di parlato emotivo (cfr. Ververidis & Kotropoulos, 2006) ha rilevato l'assenza di *corpora* di parlato emotivo mistilingue utili per uno studio cross-linguistico-culturale delle emozioni. Ciò ha reso necessario la raccolta di produzioni verbali emotive nelle quattro lingue (italiano, francese, inglese e tedesco) con riferimento alle emozioni definite da Ekman (1992) come *basic* (*happiness*, *anger*, *fear*, *sadness*, *disgust*, *surprise*).

Gli obiettivi di questa ricerca sono essenzialmente:

a. motivare e illustrare le caratteristiche del *corpus* raccolto, con particolare riferimento al protocollo di elicitazione adottato;
b. fornire i risultati della procedura di validazione. Questa operazione è assolutamente necessaria per l'attuazione delle successive fasi della ricerca, come l'analisi acustica dei campioni raccolti per la caratterizzazione delle emozioni nelle lingue in esame; la somministrazione di esperimenti percettivi nella direzione proposta da Scherer *et al.* (2001: 88) per appurare la capacità di soggetti di lingua diversa a decodificare

espressioni emotive in una lingua differente dalla propria; la stima di quanto la conoscenza della lingua possa influire sulla corretta identificazione delle emozioni presentate ecc..

Con riferimento al primo obiettivo, illustreremo nei dettagli il protocollo di elicitazione costituito da tre fasi distinte e tra loro conseguenti ispirate al 'paradigma degli scenari' e al 'contenuto standard' di Anolli *et al.* (2008a, 2008b), Anolli & Ciceri (1992) e Scherer *et al.* (1991) con la raccolta di produzioni da parte sia di soggetti *naif* che di *attori*.

Con riferimento al secondo obiettivo, invece, presenteremo i risultati della fase di validazione del corpus effettuata con due test, rispettivamente di identificazione, attraverso la verifica di eventuali differenze di riconoscimento da parte dei soggetti ascoltatori coinvolti nelle registrazioni con riferimento a ciascuna delle tre fasi del protocollo di elicitazione; e di rappresentatività delle produzioni emotive, per valutare il contributo, in termini di materiale utile e di riconoscibilità delle produzioni, da parte di soggetti *naif* da un lato e di *attori* dall'altro.

2. INTRODUZIONE

L'abilità degli esseri umani a riconoscere emozioni a partire da espressioni facciali è un dato di fatto oramai comunemente accettato dalla stragrande maggioranza degli studiosi, quanto il fatto che le stesse esistano e vengano riconosciute a livello cross-culturale.

Sulla scia del successo degli studi sulle espressioni facciali, lo studio delle emozioni espresse attraverso la voce ha subito negli ultimi anni un rinnovato interesse in diversi ambiti di ricerca e numerosi sono gli studi presenti in letteratura nelle singole lingue (sia per l'*encoding,* ricerca dei meccanismi e degli indici acustici interessati nella produzione del parlato emotivo, sia per il *decoding,* indagine sui processi percettivi e sulla capacità degli esseri umani a decodificare il parlato emotivo).

Il crescente interesse per le emozioni espresse attraverso la voce si desume soprattutto dall'ampio numero di pubblicazioni (tra cui l'istituzione di giornali e riviste dedicate) e dalla nascita di specifiche organizzazioni o reti di ricerca e cooperazione (come ad es. HUMAINE, ISER, ecc.) su tale ambito di studio.

Gli obiettivi perseguiti sono anch'essi assai diversi e spaziano dalla mera indagine tendente a chiarire le regole sottese alla trasmissione e alla caratterizzazione delle emozioni attraverso la voce, all'implementazione di parlato emotivo in sistemi di sintesi vocale ecc.

Se da un lato si rilevano innumerevoli studi sul parlato emotivo nelle singole lingue, dall'altro gli studi di tipo cross-linguistico e cross-culturale risultano essere assai sparuti.[1] I motivi non sono affatto legati, come già detto, all'assenza di interesse per l'argomento, ma sono prevalentemente imputabili alla difficoltà che lo studio delle emozioni vocali impone, difficoltà ulteriormente esacerbate nello studio di tipo cross-linguistico-culturale.

[1] Per gli studi cross-linguistici volti all'*encoding* si vedano ad esempio Anolli *et al.,* 2008a, 2008b; Braun & Oba, 2007; Kori & Magno Caldognetto, 2003; Piôt, 1999; per quelli volti al *decoding* si vedano, invece, gli studi citati e riportati in Pavlenko, 2005, a cui vanno aggiunti Pell *et al.* 2009; Sawamura *et al.*, 2007; Shochi *et al.*, 2007; Droomey *et al.*, 2005; Tickle, 1999 e 2000; Magno Caldognetto & Kori, 1983. Ancora meno sono quelli che hanno affrontato entrambi gli aspetti di *encoding* e *decoding* (ad es. Chung, 1999 e 2000; Abelin & Allwood, 2000 e 2002; Breitenstein *et al.*, 2001; Thompson & Balkwill, 2006).

Tuttavia, dopo un forte impulso iniziale dato dagli studi sulle espressioni facciali a cui è seguito un periodo di apparente stagnazione, l'argomento sta riprendendo nuovamente piede producendo un rinnovato interesse in diverse aree di studio.

Dai risultati degli studi sino ad oggi condotti emerge infatti come le emozioni vocali siano, alla pari di quelle facciali, riconosciute cross-linguisticamente e cross-culturalmente, con risultati nettamente al di sopra della semplice casualità. Da una meta-analisi condotta da Laukka (2004) sugli studi presenti in letteratura riguardo alle emozioni vocali, allo scopo di verificare se le emozioni siano riconosciute cross-culturalmente e se vi siano specifici *patterns* acustici della voce per categorie discrete di emozioni, emerge come: a) l'accuratezza di riconoscimento è superiore a quella data dalla casualità per categorie di emozioni più ampie; b) il *decoding* cross-culturale è mediamente inferiore al *decoding* intra-culturale del 7%; c) esistono specifici *patterns* acustici nella voce delle emozioni che vengono utilizzati per comunicare emozioni discrete.

Ciononostante, se da una parte Scherer, Banse & Wallbott (2001: 78) affermano che "it seems reasonable to assume that the recognition of vocal emotion expressions might work across language and culture boundaries", più avanti gli stessi autori, sottolineano come "[…] encoders and decoders from several different countries would need to be studied, allowing the construction of an encoder-decoder-emotion matrix and to test whether decoders from the countries involved would recognize emotion portrayals by encoders from their own countries most accurately" (Scherer, Banse & Wallbott 2001: 88).

È proprio nel quadro qui descritto da Scherer, Banse & Wallbott che si inserisce il presente lavoro.

3. LO STUDIO DELLE EMOZIONI VOCALI: QUESTIONI METODOLOGICHE

Prima di affrontare la tematica principale all'oggetto del lavoro qui presentato, occorre soffermarsi un attimo su alcune questioni di ordine metodologico nello studio delle emozioni vocali. Non è infatti un caso che, alle ricerche sul parlato emotivo si aggiungano, oltre ad una serie di problematiche simili a quelle delle espressioni facciali, tutta una serie di altre difficoltà ulteriormente complicate in termini metodologici nel caso di ricerche condotte a livello cross-linguistico.

Una delle più grandi e significative differenze tra la ricerca sulle emozioni trasmesse attraverso le espressioni facciali e quella sulle emozioni comunicate tramite espressioni verbali è data soprattutto dalla diversità di approccio che lo studio in questione richiede, differenze che Pell *et al.* (2009: 108) sintetizzano nel seguente modo: "Contrary to research on the face, researchers interested in the voice cannot present a valid 'snapshot' which represents the vocal attributes of an emotion." Si tratta di una differenza per nulla banale. Come fanno notare Poggi & Magno Caldognetto (2004), il compito di chi si è occupato di parlato emotivo è risultato, e tutt'ora lo è, ancor più complicato a causa della grande variabilità dei dati raccolti, siano essi di tipo intra-linguistico che di tipo inter-linguistico, variabilità dovuta alle limitazioni metodologiche (tipo di parlato spontaneo *vs.* parlato di laboratorio, letto o recitato); alla scelta dei parlanti (*attori* vs. parlanti *naif*); al metodo di elicitazione delle emozioni (fotografie, scenari, etichette linguistiche ecc.) e, di non minore importanza, alla scelta del materiale linguistico prodotto (vocali, sillabe, interiezioni, stringhe fonologiche non-sense ecc.). Per tali ragioni è quindi opportuno mettere in evidenza alcune questioni di ordine metodologico con le relative problematiche connesse,

sulla base delle quali è stato successivamente possibile adottare determinate strategie volte alla creazione del corpus emotivo mistilingue che presenteremo.

3.1 Limitazioni di tipo etico e morale

Contrariamente all'opinione comune, anche nella ricerca scientifica sulle emozioni esistono limiti di tipo etico e morale. In tal senso sono infatti numerosi i ricercatori che sino ad oggi si sono adoperati per riuscire a superare i limiti imposti da questioni di natura etica, qualunque fosse la motivazione dietro allo studio delle emozioni (Rottenberg *et al.*, 2007).

La questione 'etica' non è quindi cosa di poco conto. Nel mondo occidentale, ma oramai ovunque, la ricerca scientifica prevede che ci si conformi e che si rispettino specifiche norme di tipo etico e morale. Tuttavia, come fanno rilevare Niedenthal *et al.* (2006), le limitazioni etiche e morali non sono sempre state così evidenti come dimostra un esperimento condotto negli anni '50 del secolo scorso da Ax (1953) e che oggi sarebbe assolutamente improponibile.

Il ricercatore non può ricorrere a tutti gli stratagemmi che gli vengono in mente e non può, in ogni caso, indurre nei soggetti sperimentali emozioni o sensazioni oltre quelle che ci si aspetterebbe nella vita quotidiana da parte di soggetti non affetti da disturbi o psicopatologie a carico della sfera affettiva. Allo stesso modo, le emozioni indotte sperimentalmente dovrebbero poter essere suscitate facendo ricorso a stimoli o situazioni comunemente riscontrabili nella vita di tutti i giorni e l'induzione di emozioni fortemente dolorose viene spesso limitata o addirittura rifiutata. Anche quando il ricorso all'induzione di emozioni negative dovesse rendersi assolutamente necessario, le stesse emozioni indotte dovranno essere transitorie, ovvero devono estinguersi non appena il soggetto abbia abbandonato il laboratorio, senza lasciare segni nel soggetto utilizzato nella sperimentazione.

3.2 Catturare le emozioni

Disporre di procedure di elicitazione valide e funzionali agli scopi della ricerca che si intende intraprendere, e da cui ricavare il materiale oggetto di studio e di analisi, rappresenta, anche in virtù delle limitazioni sopra esposte, un obiettivo di primaria importanza a cui molti ricercatori e studiosi interessati alle emozioni nei vari ambiti, si sono impegnati nel tentativo di mettere a punto procedure che fossero quanto più possibile controllate e replicabili.

Sebbene il recente *Handbook of Emotion Elicitation and Assessment*, curato da James A. Coan e John J.B. Allen vanti una ricca ed esaustiva raccolta delle metodologie più comuni per l'elicitazione e l'induzione di stati affettivi/emotivi in laboratorio,[2] resta indubbio il fatto che riuscire a catturare le emozioni non sia cosa semplice.

Le tecniche di raccolta sono spesso ed esclusivamente di tipo sperimentale, e su queste non esiste nemmeno grande consenso. Da più parti si assiste infatti ad un'impressionante proliferazione di nuove tecniche di elicitazione che, spesso e volentieri, non trovano sufficiente spazio per i dovuti approfondimenti, e che per questo motivo vengono relegate a sommarie descrizioni.

Ciò risulta ancor più vero se si considera che, da un lato, quello delle scienze affettive rappresenta oggi un dominio di ricerca non più associato alla sola psicologia, ma investe e chiama in causa una serie di altre discipline, all'interno delle quali, ciascuno dei suoi

[2] Il volume dedica, infatti, sotto il nome di *Emotion elicitation*, un'intera sezione di 10 capitoli a quelle che sono le tecniche di elicitazione e raccolta più diffuse e conosciute.

appartenenti tenta di dare il proprio contributo attraverso le proprie tecniche e le proprie conoscenze; dall'altro, ed è questo l'aspetto che appassiona e intriga di più, le emozioni, intese nell'accezione più generica del termine, investono e influenzano ogni livello di indagine chiamandone in causa altri. Lo studio delle emozioni come tale, oltre a richiedere specifiche competenze che abbracciano più di una disciplina, richiede disegni sperimentali e strategie ben definite, e tecniche di indagine e strumentazioni anche molto sofisticate dove la collaborazione tra ricercatori e laboratori diventa essenziale, se non addirittura inevitabile.

Nell'ambito dello studio delle emozioni espresse attraverso la voce, le strategie e le tecniche di raccolta del parlato emotivo si differenziano tra loro per tipo di approccio utilizzato e numero di emozioni che sono in grado di indurre.

La scelta di una metodologia piuttosto che l'altra è strettamente connessa alle finalità e al disegno sperimentale della ricerca che si intende intraprendere, con la consapevolezza che la stessa sarà caratterizzata da vantaggi e svantaggi che vanno tenuti in debita considerazione. Dal momento che non esistono procedure standardizzate per la raccolta di parlato emotivamente connotato, anche il tentativo di una suddivisione di quelle che possono essere le tecniche utilizzate diventa assai difficile, se non addirittura fuorviante, proprio perché molte di queste si compenetrano e si completano a vicenda (cfr. Rottenberg *et al.,* 2007).

Un altro problema a cui alcuni studiosi si sono dedicati, riguarda la possibilità di stabilire quanto effettivamente l'utilizzo di una procedura di elicitazione sia efficace. Vi sono innumerevoli studi che evidenziano differenze sostanziali tra le varie metodologie in riferimento alla loro efficacia: si va infatti da un 87% di riuscita per tecniche che fanno uso della musica come strumento di elicitazione, per piombare ad un bassissimo 15% per tecniche che fanno uso di suggestione ipnotica.[3] In una meta-analisi condotta su 138 studi da Westermann, Spies, Stahl, Hesse (1996), invece, la presentazione di film o storie si è rivelata essere la tecnica più efficace nell'induzione di stati affettivi sia positivi che negativi.

Sebbene una chiara suddivisione delle tecniche utilizzate sia impossibile, in letteratura sono presenti procedure che producono:

- parlato emotivo 'indotto' sperimentalmente con l'ausilio di *Mood Induction Procedures* (MIPs) che utilizzano particolari procedure di laboratorio a cui il soggetto viene sottoposto, e che producono materiale emotivo che si potrebbe definire 'semi-naturale';
- parlato emotivo 'naturale', e quindi spontaneo e autentico, ricavato da situazioni di vita quotidiana o frutto di particolari interazioni uomo-uomo o anche uomo-macchina;[4]
- parlato emotivo 'simulato', o recitato, anche qui secondo schemi e modalità predeterminate.[5]

[3] Cfr. Eich *et al.* (2007: 125).

[4] Sulla possibilità di utilizzo di parlato di questo tipo sussistono questioni legate alla privacy che rientrano in quelle che sono state definite nel precedente § come "Limitazioni di tipo etico e morale". Nonostante ciò si assiste, allo stato attuale, ad un crescente ricorso a parlato emotivo 'reale' proprio per le opportunità e per le possibilità che questo tipo di materiale offre in applicazioni commerciali *real-time*.

Tenendo presenti i limiti relativi alla raccolta di materiale emotivo descritti in questo paragrafo, nel prosieguo si tenterà di fornire una visione il più possibile esauriente delle tecniche più comunemente utilizzate per la raccolta di produzioni emotive, sottolineando il fatto che molte di esse vengono fatte interagire rendendo in alcuni casi assai difficile la tripartizione a cui più sopra si accennava.

3.2.1 Emozioni indotte: le Mood Induction Procedures (MIPs)

Le MIPs sono per lo più tecniche o strategie di laboratorio, "whose aim is to provoke in an individual a transitory emotional state in a non natural situation and in a controlled manner; the mood induced tries to be specific and pretends to be an experimental analogue of the mood that would happen in a certain natural situation".[6]

Con riferimento alle MIPs, Eich *et al.* (2007: 125) sottolineano come tecniche sperimentali di questo tipo siano state messe a punto per prime e come siano in forte aumento e diventino sempre più popolari, proponendo una *wishlist* di sei punti a cui una MIP dovrebbe rispondere e che qui di seguito viene riprodotta:

Desirable attributes of a mood induction technique
• Technique has a high rate of success in altering participants' moods in predictable ways.
• Technique allows for individual differences in time taken to develop a particular mood.
• Induced moods are strong or intense.
• Induced moods are stable over time and across tasks.
• Induced moods seem real or authentic to the participants.
• One and the same mood can be reliably induced on more than one occasion.

Tabella 1: *Wishlist* degli attributi che una tecnica di *mood induction* dovrebbe possedere, proposta (e adattata) da Eich *et al.* (2007: 125), tab. 8.1

Esempi di induzione di emozioni prevedono l'uso di sequenze video o di film di cui viene fornita una accurata descrizione in Rottenberg *et al.* (2007). Altri propongono come MIPs particolari tecniche di realtà virtuale, come ad esempio Baños *et al.* (2006) che utilizzano un parco virtuale che cambia in relazione all'umore da indurre. Altri ancora hanno fatto uso di brani musicali, tecnica utilizzata per la prima volta da Sutherland, Newman & Rachman (1982), chiedendo a dei soggetti di utilizzare i brani musicali presentati come mezzo per entrare in uno stato d'animo, avvertendoli, allo stesso tempo, del fatto che la musica da sola non è in grado di indurre in loro lo stato d'animo in modo automatico, ma che dovevano fare ricorso a proprie strategie per raggiungerlo. Alcuni si sono avvalsi di specifiche tecniche di induzione per valutarne gli effetti sulla manipolazione della voce, come ad esempio Johnstone *et al.* (2005) che hanno utilizzato un videogame per l'induzione di espressioni emotive naturali in laboratorio. Ulteriori tecniche per indurre determinate emozioni, oggi sicuramente meno attuabili per le ragioni a cui più sopra si è accennato, prevedono l'utilizzo di droghe come in Helfrich *et al.* (1984) in cui sono stati

[5] La presente tassonomia, relativa ai tre tipi di parlato generalmente raccolto e riscontrabile negli studi sulle emozioni vocali, fa riferimento a quella riportata da Scherer (2003: 231 ss.).

[6] Cfr. García-Palacios & Baños (1999: 16), *op. cit.* in Baños *et al.* (2006).

studiati gli effetti di antidepressivi sui diversi parametri vocali per un periodo di diverse ore.

I vantaggi derivanti dall'utilizzo delle MIPs sono dati dal fatto che si riescono ad ottenere emozioni pressoché spontanee con un controllo assoluto della situazione di laboratorio appositamente creata; al contrario, nella maggior parte dei casi le emozioni prodotte risultano però difficili da identificare o da etichettare per l'elevata variabilità che caratterizza le produzioni da un soggetto all'altro. Tuttavia, come già accennato, accanto alla possibilità che queste tecniche offrono nell'indurre emozioni 'reali', a seconda dei casi sussistono limitazioni di tipo etico che difficilmente sono superabili.

Pro	Contro
Emozioni spontanee e 'reali'	Difficoltà di identificazione e/o etichettatura
Elevata ecologicità	Elevata variabilità
Controllo sulle procedure di induzione	Limitazioni di tipo etico

Tabella 2: Tabella riassuntiva degli elementi a favore o a sfavore delle MIPs

A questa tipologia appartiene anche il metodo cosiddetto 'Velten' (Velten, 1968), una tecnica di MIP per l'induzione di stati affettivi positivi e/o negativi attraverso il comportamento verbale. Nella sua versione originaria[7] la procedura prevede che il metodo venga somministrato oralmente e individualmente ai partecipanti a cui viene chiesto di leggere una lunga lista di affermazioni emotivamente caratterizzate, dapprima a se stessi e poi ad alta voce. Le affermazioni somministrate vanno da uno stato inizialmente neutro per arrivare a connotazioni emotive positive o negative a seconda dei casi presentati[8], e sono così in grado di produrre cambiamenti nell'umore del soggetto che le legge.

In letteratura il metodo si ritrova in diverse varianti e con differenti gradi di riuscita. Proprio in riferimento al grado di efficacia di tale metodo e di altri metodi simili si sono pronunciati in molti, tra cui Gerrards-Hesse *et al.* (1994), Westermann *et al.* (1996), ma anche Clark (1983). Secondo quest'ultimo, ben un terzo, se non addirittura la metà dei partecipanti, manifesta poco o nessun cambiamento di umore in risposta a tale metodo. Dello stesso avviso è anche Kenealy (1986: 331), che in uno studio condotto a partire dalle risultanze sperimentali della Velten MIP su 46 esperimenti, giunge alla conclusione che: "the findings relating to the Velten procedure […] are inconsistent and equivocal".

3.2.2 Emozioni 'autentiche' o naturali

Un'altra strada molto praticata nello studio delle emozioni espresse per mezzo della voce prevede l'utilizzo di parlato emotivo 'reale' o 'autentico' che, come detto, risulta di grande interesse per le opportunità e per le possibilità che questo tipo di materiale offre in applicazioni commerciali real-time.

Riguardo questo aspetto Campbell (2000) mette però in evidenza come, ad esempio, i database utilizzati in ambito di *speech technology* facciano per lo più uso di parlato

[7] Ne esiste anche una versione rivista adottata da Seibert & Ellis (1991) in cui, rispetto alla versione di Velten (1968), gli item della procedura contengono espressioni di uso familiare e di uso quotidiano tipico degli studenti, non contengono riferimenti a stati d'animo di tipo somatico e si presenta con solo 25 produzioni invece di 60.

[8] Le affermazioni, dell'ordine di ca. 60 frasi, vengono presentate su singole schede (*cards*).

prodotto da attori, rischiando di rappresentare in modo falsato le caratteristiche del parlato emotivo reale con conseguenze a danno di coloro che sono i fruitori delle applicazioni create e delle soluzioni tecnologiche che su tali materiali vengono modellate: "If a computer speech synthesizer were to emulate such speaking-style characteristics successfully, then it may be liable to misrepresent the intentions of its user. For example, if a disabled person for whom speech synthesis is the sole means of verbal expression, was to use the synthesized voice to express genuinely-felt pleasure (or anger), then the listener might be able to 'hear' that the voice was only expressing acted pleasure (or anger), and is liable to (mis-)respond accordingly."

Tuttavia, seguendo il consiglio di Campbell (2000), sebbene l'utilizzo di emozioni catturate in contesti o situazioni di vita reale sia da preferire, in quanto caratterizzate da una elevatissima validità ecologica, come fanno notare Scherer, Johnstone & Klassmeyer (2003: 436), anche per questa tipologia di parlato emotivo sussistono una serie di problematiche connesse: l'utilizzo di parlato emotivo appartenente a questa tipologia viene per lo più criticato per il fatto che esso avviene in situazioni di vita reale dove diventa pressoché impossibile definire con esattezza quale emozione venga esperita in quel preciso momento dal soggetto che viene registrato; a ciò si aggiunga anche che spesso le produzioni raccolte sono di brevissima durata e di bassissima qualità per l'impossibilità di controllare le modalità di acquisizione rispettando le più basilari tecniche di registrazione, registrazioni ulteriormente affette da eventi provenienti dall'ambiente circostante.

3.2.3 Emozioni simulate o 'posate'

Solitamente, quando si parla di emozioni simulate, o ancora posate o recitate, la nostra mente chiama inevitabilmente in causa gli attori.

L'utilizzo di attori per la raccolta di parlato emotivamente caratterizzato risale solitamente a quelli che vengono conosciuti come *Fairbanks studies* sul finire degli anni '30, considerati da molti come i primi studi sistematici di tipo sperimentale sul parlato emotivo.[9]

In questo caso il metodo di cui solitamente ci si avvale è quello 'Stanislavskij'. Si tratta di una tecnica di insegnamento della recitazione messa a punto da Kostantin Sergeevič Stanislavskij nei primi anni del '900. Tale metodo si basa sulla esternazione di emozioni interiori attraverso la loro interpretazione e rielaborazione a livello intimo e a livello di esperienze del vissuto personale dell'attore stesso. Si tratta, in definitiva, di quello che può essere considerato un processo di auto-induzione di stati emotivi a partire dalla propria esperienza di vita.

L'utilizzo di questo metodo per la raccolta di parlato emotivamente caratterizzato a fini di ricerca pone, però, una serie di questioni. Tra quelli che possono essere definiti vantaggi vi sono sicuramente la possibilità di ottenere materiale emotivo anche molto 'intenso' che ricade positivamente sulla possibilità che le produzioni prodotte vengano correttamente identificate o riconosciute da gruppi di ascoltatori in prove di tipo percettivo. Questa 'intensità' delle produzioni raccolte si scontra, tuttavia, con quella che è rappresentata dalla

[9] Vedi Fairbanks, G., Hoaglin, L. W. (1941), An experimental study of the durational characteristics of the voice during the expression of emotion, *Speech Monograph*, 8, 85-91 e Fairbanks, G., Pronovost, W. (1939), An experimental study of the pitch characteristics of the voice during the expression of emotion, *Speech Monograph*, 6, 87-104, *op. cit.* in Schröder (2004: 43).

realtà e dalla vita quotidiana: in che misura, in condizioni normali come quelle della vita di tutti i giorni, si assiste a realizzazioni così intense? L'attore, nel rivivere e riprodurre determinate emozioni, tende forse a rappresentare quella che dovrebbe essere la realtà in modo troppo esagerato o stereotipato: questo porterebbe sicuramente ad una maggiore riconoscibilità della produzione emotiva, ma porterebbe allo stesso tempo a considerare innaturale o poco realistica l'emozione codificata. Un altro aspetto a sfavore di questa tecnica in contesti scientifici quali quelli previsti dalla ricerca, è dato dal fatto che non vi è assolutamente alcun controllo sulle esperienze (soggettive) richiamate dal singolo attore nella realizzazione delle emozioni richieste: ciò causa inevitabili ripercussioni su quello che è il significato che una determinata emozione assume da un soggetto all'altro e che è strettamente legato alla situazione che l'ha generata. Infine, se da un lato il ricorso a questa tecnica consente di ridurre al minimo il dispendio di tempo per la raccolta del corpus, in quanto si tratta di produzioni in un certo senso mirate ai fini della ricerca che si intende condurre, dall'altro si pone il problema del reperimento di attori professionisti che si prestino a simili operazioni e che raramente, fatte salve alcune eccezioni, prestano la loro opera gratuitamente.

Nonostante i vantaggi e i limiti appena esposti, sono tanti coloro che fino ad oggi hanno fatto ricorso a questa tecnica. Burkhardt *et al.* (2005), ad esempio, hanno fatto uso di questa metodologia per la creazione di un database di parlato emotivo recitato in lingua tedesca contenente 10 frasi prodotte in sei stati emotivi target da 10 attori, fornendo agli attori coinvolti semplici istruzioni orali e facendo esplicito affidamento sulle loro abilità ad auto-indursi determinate emozioni. In modo analogo ha operato Iadarola (2009) nella raccolta di un database di parlato emotivo per l'italiano in cui i soggetti coinvolti erano tutti attori.

Va tuttavia tenuto presente che gli studi basati esclusivamente sulle linee guida imposte dal metodo 'Stanislavskji' sono di numero decisamente inferiore di quelli che, per non incorrere nelle critiche rivolte a questa tecnica, hanno fatto ricorso ad attori con accorgimenti e strategie metodologiche diverse volte proprio a superare, o quanto meno a ridurre, alcune significative limitazioni.[10] È infatti a questo proposito che Enos & Hirschberg (2006) propongono due particolari approcci per l'elicitazione di parlato emotivo con attori, proprio per far fronte alle difficoltà che pone l'utilizzo del parlato recitato nell'ambito delle ricerche sulle emozioni.[11]

[10] Il problema non è ovviamente legato al metodo in quanto tale, ma al suo utilizzo in questo specifico ambito. Uno degli elementi che suscita maggiore critica è dato dal fatto che il più delle volte all'attore venga semplicemente chiesto di pronunciare un testo 'X' con una emozione 'Y' senza istruzioni aggiuntive, lasciando appunto troppo spazio alla 'interpretazione' soggettiva.

[11] Vedi Enos & Hirschberg (2006) per maggiori dettagli sui due approcci proposti.

3.2.4 Gli scenari: una soluzione 'ibrida'

Si è più sopra fatto accenno all'utilizzo di scenari, o *scenario approach*, per la raccolta di parlato emotivo.[12] Con il termine scenario ci si riferisce ad una descrizione sintetica di un evento o di una serie di azioni e di eventi, la cui organizzazione e/o composizione segue quelle che sono le tecniche narrative. Come già evidenziato per molte altre tecniche, anche questo tipo di approccio si presenta in diverse varianti.

Nella variante utilizzata da Anolli *et al.* (2008a, 2008b) e da Anolli & Ciceri (1992),[13] in cui gli scenari sono stati somministrati a soggetti *naif*, si prevede la presenza e la definizione di due elementi fondamentali: a) una situazione che fornisce le condizioni caratteristiche della specifica emozione; b) una serie di risposte standard alla situazione. A partire da questi due elementi in ciascun testo sono state inserite in modo sistematico informazioni:

- 'stimolo pertinenti', ovvero informazioni su stimoli esterni in grado di attivare una data emozione, stimoli allo stesso tempo coerenti con l'emozione in questione;
- 'risposta pertinenti', ovvero una serie di informazioni riguardo le risposte emotive più consone al contesto e comprendente risposte verbali, esclamazioni, reazioni motorie ecc.;
- 'di significato', ovvero informazioni che definiscono il significato dello stimolo emotigeno e delle risposte da esso derivanti con riferimento all'*appraisal*.[14]

Una rappresentazione schematica della struttura di uno scenario per come proposto nell'approccio appena descritto, potrebbe essere data dalla seguente figura:

[12] L'approccio degli scenari potrebbe essere considerato come metodo di induzione di emozioni e sarebbe quindi dovuto essere trattato nel paragrafo dedicato al parlato emotivo indotto. Tuttavia, come si vedrà nel prosieguo, questo approccio lascia un certo grado di libertà ai soggetti coinvolti nelle operazioni di raccolta di parlato emotivo, e più che 'indurre' determinate emozioni si parla in questo caso di 'elicitare' le emozioni facendole, in un certo qual modo, emergere con un certo grado di naturalezza consentendo, allo stesso tempo, di controllare una serie di variabili a cui si accennerà più avanti.

[13] Gli autori seguono le proposte di precedenti studiosi tra cui De Sousa (1987), Rosenthal *et al.* (1979) utilizzando anche una versione adattata del "paradigma del contenuto standard" di Davitz (1964) consistente nella presenza di una frase dal contenuto standard in tutti gli scenari proposti. Nella variante presentata, per la lingua italiana la frase in questione era "non è possibile, non ora" (cfr . Anolli *et al.* 2008a: 8). Per una visione di quelli che sono i testi degli scenari utilizzati si rimanda ad Anolli *et al.* (2008b) e Anolli & Ciceri (1992).

[14] Cfr. Anolli *et al.* (2008a: 10). Il termine *appraisal* si riferisce ai meccanismi di valutazione cognitiva attraverso i quali vengono organizzate le esperienze emotive.

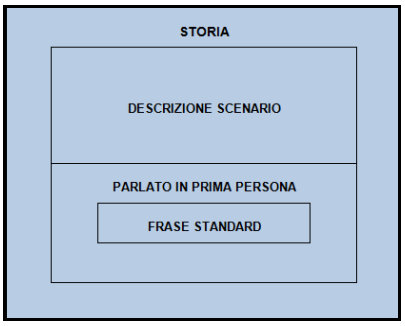


Figura 1: Rappresentazione schematica del costrutto di uno scenario o *frame story* nei termini proposti da Anolli *et al.* (2008a, 2008b) e Anolli & Ciceri (1992)

Nell'altra variante di questo approccio, utilizzata sino ad oggi da un numero maggiore di studiosi, è invece prevista la somministrazione, in questo caso, ad attori professionisti, di scenari o situazioni molto più sintetici con specifiche istruzioni.[15] Nel caso di Scherer *et al.* (1991), gli scenari sono molto sintetici, come ad esempio quelli utilizzati per *anger:* "1) The director is again late for rehearsal and we have to work until very late at night. Once again I have to cancel a date. 2) I have sublet my apartment for a period of several months. Upon my return my place is in a real mess and the person did not keep to a single agreement".[16]

In questo caso agli attori è stato chiesto di leggere gli scenari e di immaginare di vivere o di esperire la situazione descritta e successivamente di pronunciare una frase standard[17] nello stesso modo in cui l'avrebbero pronunciata se si fossero trovati in quella situazione.

Il concetto di 'ibrido', a cui si è fatto qui riferimento, consiste nella possibilità di indurre e allo stesso tempo di far simulare (entro certi limiti) su una base comune (data dagli scenari) le emozioni desiderate utilizzando sia soggetti *naif* che *attori*. Tale approccio consentirebbe di aggirare alcune delle limitazioni di cui ai precedenti paragrafi, sebbene, come già evidenziato per le precedenti modalità di raccolta di parlato emotivo, anche in questo caso non mancano vantaggi e svantaggi che vengono riportati nella tabella 3:

[15] Considerando quindi quest'altro aspetto l'argomento qui trattato avrebbe dovuto trovare spazio nel paragrafo dedicato alle emozioni simulate o 'posate'.

[16] Cfr. Scherer *et al.* (1991: 147). Gli scenari proposti dagli autori sono stati ricavati da studi di tipo inter-culturale precedentemente condotti in cinque continenti sull'esperienza delle emozioni in cui sono state raccolte situazioni rappresentative dell'elicitazione delle emozioni su un campione di oltre 3.000 intervistati (cfr. Wallbott & Scherer, 1986).

[17] In questo caso sono state usate "speech-like but meaningless sentence(s)" create da un fonetista sulla base di elementi fonemici di varie lingue europee: *"Hat sundig pron you venzy"* e *"Fee gott laish jonkill gosterr"* (Wallbott & Scherer, 1986: 126). La scelta di questo tipo di frasi era stata fatta in vista di un successivo studio a livello cross-culturale.

Pro	Contro
Controllo della situazione elicitante e delle emozioni	Gli scenari vanno creati o adattati allo scopo
Facilità di attuazione	Molti soggetti da registrare
Abbondanza di materiale	Richiede molto tempo

Tabella 3: Tabella riassuntiva dei vantaggi e degli svantaggi relativi ad una soluzione di raccolta 'ibrida' a cavallo tra emozioni simulate ed emozioni indotte

3.3 Scelta del materiale linguistico

La scelta del materiale linguistico o del tipo di parlato da utilizzare nelle ricerche sul parlato emotivo rappresenta, da questo punto di vista, un elemento cruciale che merita la dovuta considerazione.

Strettamente connessa al materiale linguistico scelto risulta infatti la possibilità di effettuare determinate analisi di tipo acustico, o ancora, di tipo percettivo.

Sebbene le possibilità di scelta siano pressoché infinite (considerando le caratteristiche di ricorsività della lingua), queste sono sostanzialmente riconducibili a frasi (come in Anolli & Ciceri, 1992; Walbott & Scherer, 1986), frasi non-sense (Scherer *et al.* 2001), ma anche parole isolate (Piôt, 1999), cifre numeriche, sequenze 'VCV (come ad esempio /'aba, 'ava/ in Magno Caldognetto *et al.*, 2005), vocali isolate (Magno Caldognetto & Kori, 1986) o ancora affect bursts (Schröder, 2003).

Come è facile intuire, nell'elenco delle possibilità appena riportato non solo la quantità di materiale utile diminuisce a seconda della tipologia di materiale utilizzato, ma di contro aumentano le difficoltà delle analisi che sul materiale raccolto si intendono effettuare.

3.4 Le etichette verbali emozionali

Nello studio delle emozioni vocali anche le 'semplici' etichette verbali, come ad esempio gioia, tristezza e collera, hanno la loro importanza.

Uno degli aspetti spesso trascurati riguarda il fatto che una singola etichetta emozionale possa rimandare o riferirsi a più di uno stato affettivo: non esiste la collera, ma la collera fredda e la collera calda.

Nel caso vengano condotte ricerche all'interno di una singola lingua le difficoltà maggiori sono sostanzialmente legate al fatto che soggetti diversi possono attribuire diverso significato emotivo a specifiche etichette verbali, come nel caso già richiamato. O ancora, come evidenziano Scherer *et al.* (2001: 79), soggetti diversi possono usare situazioni elicitanti diverse per una stessa emozione.

La questione si complica ulteriormente nel caso in cui la ricerca o lo studio preveda che queste debbano essere tradotte in un'altra lingua come nel caso di studi di tipo cross-linguistico. Prima di affrontare tali questioni più da vicino, si dia uno sguardo alla seguente tabella che rappresenta il punto di partenza nel caso si voglia affrontare uno studio di tipo cross-linguistico:

Italiano	Français	English	Deutsch
collera (calda)	colère (chaude)	anger (hot)	Ärger (warmes)
gioia	joie	joy	Freude
paura	peur	fear	Angst
disgusto	dégoût	disgust	Ekel
tristezza	tristesse	sadness	Traurigkeit
sorpresa	surprise	surprise	Überraschung
neutro	neutre	neutral	Neutral

Tabella 4: Etichette verbali relative a sei stati affettivi in quattro lingue europee, ad. da: K. R. Scherer (editor) (1988), Labels describing affective states in five major languages, *Facets of emotion: Recent research*, Hillsdale, NJ: Erlbaum, 241-243

Da quanto riportato si evince una corrispondenza uno a uno dei termini riportati in appendice nel volume curato da Scherer (*op. cit*). Tuttavia, alle problematiche già individuate all'interno di ogni singola lingua, si aggiungono qui ulteriori criticità. Come evidenziato da Averill (2004: 578), le etichette verbali differiscono da lingua a lingua e non sempre il rapporto tra i termini è univoco. Dove una lingua, espressione di una cultura, e in quanto tale interpretazione della realtà, contempla un solo termine per uno o più stati affettivi, in un'altra viene operata una netta distinzione attribuendo determinate etichette emozionali a determinati stati affettivi: ad es., contrariamente all'italiano che non distingue ulteriormente il concetto di tristezza, identificato da una sola etichetta, il tedesco prevede un'etichetta *Traurigkeit* per la tristezza comunemente intesa, mentre si riferisce a *Trauer* nel caso di tristezza legata ad eventi luttuosi. A ciò si aggiunga, inoltre, il fatto che una stessa etichetta emozionale può avere una sola valenza in una lingua (senza distinzione tra stato positivo o negativo), mentre può avere tale distinzione in un'altra lingua. Härtel & Härtel (2005: 685-686) fanno infatti notare come, ad esempio, "the word 'surprise' in English does not necessarily have a positive or negative connotation, whereas the German word 'Überraschung' probably has more of a positive than a negative connotation".[18]

3.5 Corpora o risorse di parlato emotivo

I *corpora*, o database in generale, sono raccolte più o meno ampie di parlato, di produzioni verbali (scritte o orali) raccolte a vario titolo, che rientrano in quelle che comunemente vengono identificate con il termine di 'risorse linguistiche'. In questo contesto con il termine di *corpora* ci si riferisce esclusivamente a risorse linguistiche di parlato emotivo tralasciando, volutamente, qualsiasi altra tipologia di risorsa.

Grazie all'oramai inarrestabile innovazione tecnologica si assiste, allo stato attuale, ad un fiorente mercato attorno a tali raccolte, che vengono commissionate e progettate da vari soggetti o enti per gli scopi più svariati, soprattutto per applicazioni in ambito commerciale. Esistono enti o istituzioni come la *European Language Resources Association* (ELRA) la cui missione è quella di promuovere le risorse linguistiche nel settore dell'*Human Language Technology* (HLT) e di valutare le tecnologie di ingegneria della lingua, per

[18] Per una rassegna più completa ed esauriente si rimanda il lettore all'opera di Pavlenko A. (2005), *Emotions and Multilingualism*, New York: Cambridge University Press, in cui tali aspetti vengono chiariti e messi in relazione con il supporto di numerosi esempi.

conto della quale opera un'altra agenzia denominata ELDA (*Evaluations and Language resources Distribution Agency*) che si occupa, nello specifico, della valutazione e della distribuzione di risorse linguistiche.[19] Per quello che in questa sede può interessare, è assai singolare il fatto che, all'interno del catalogo delle risorse accessibili dal sito web dell'agenzia, una ricerca effettuata con la parola chiave *emotion* abbia prodotto come unico risultato un rimando a sole tre raccolte di uno stesso progetto.[20]

Occorre a questo punto richiamare una distinzione operata da Douglas-Cowie *et al.* (2003) che forse può contribuire a fare luce sull'apparente carenza di risorse di parlato emotivo nei canali 'ufficiali' di distribuzione di tali risorse. Più che assistere alla presenza di veri e propri *database*, nella letteratura sulle emozioni vocali si assiste alla presenza di *dataset*, ovvero di raccolte di dimensioni relativamente ridotte, generalmente costituite per studiare singoli aspetti del parlato emotivo.

Come fanno notare Douglas-Cowie *et al.* (2003: 33) la ricerca sul parlato e sulle emozioni, "is moving from a period of exploratory research into one where there is a prospect of substantial applications, notably in human–computer interaction."

Il fatto appena riportato non è poi del tutto banale in quanto, come già detto, le raccolte di parlato emotivo attualmente esistenti sono il frutto di vari esperimenti e di varie ricerche sul parlato emotivo condotte negli anni. Quello delle risorse linguistiche è un tema assai caro alla comunità dello *speech* dove non mancano certo i tentativi di creare simili raccolte. È proprio sulla base delle esperienze, delle difficoltà e delle scelte operate negli studi che negli anni si sono succeduti che Douglas-Cowie *et al.* (2003) mettono in luce una serie di caratteristiche che le nuove generazioni di *database* dovranno tenere in considerazione:

1. Scopo. Un database dovrebbe considerare e includere il maggior numero di variazioni possibili, come ad esempio numero di parlanti, lingua parlata, sesso, emozioni considerate ecc., che ne consentano un utilizzo adeguato, generalizzato ed ampio. La sua progettazione e creazione dovrebbe inoltre tener conto dei possibili ambiti di utilizzo.
2. Naturalezza. Esistono diverse strategie di raccolta del parlato emotivo, strategie che danno origine a produzioni naturali, semi-naturali e simulate o posate. Sulla scelta delle tipologia di parlato più idonea prevalgono, in ogni caso, gli obiettivi della ricerca che stabilisce quale di esse sia la più adeguata.
3. Contesto. L'ascoltatore fa generalmente riferimento al contesto per comprendere il significato emotivo di determinate caratteristiche vocali. Quattro sono i tipi di contesto identificati dagli autori:
    a. semantico: nel parlato emotivo a volte ci si avvale di parole emotivamente marcate stabilendo una potenziale interazione tra contenuto e segnale vocale.
    b. strutturale: molti eventi o segnali emotivi vengono codificati in base a strutture sintattiche (accenti, *patterns* intonativi ecc.) o in base a variazioni di stile

[19] Per maggiori dettagli si rimanda a http://www.elra.info/.

[20] Il progetto è il VERBMOBIL II, un progetto a lungo termine del Ministero Federale dell'Educazione e della Ricerca (*Bundesministerium für Bildung und Forschung – BMBF*) tedesco ed ha lo scopo di fornire alla Germania una posizione di punta nelle tecnologie del linguaggio e alle sue applicazioni economiche grazie ad una cooperazione e concentrazione del maggior numero possibile di specialisti provenienti dall'industria e dalla scienza.

attraverso le caratteristiche strutturali delle frasi (lunghe o corte, ripetizioni, interruzioni ecc.).

c. intermodale: non sempre il parlato rappresenta una fonte di informazione *stand-alone*, ma spesso funge da supplemento ad altre fonti di informazione quale quelli della faccia e dei gesti.
d. temporale: il parlato naturale veicola determinati indici di variazione lineare e sequenziale nel momento in cui le emozioni si manifestano e svaniscono nel corso del tempo.

4. Descrittori. Un *database* richiede tecniche di descrizione del contenuto linguistico ed emozionale da un lato e del parlato dall'altro.
5. Accessibilità. Il valore e l'importanza di un database aumenta ed è tale solo se esso è accessibile all'intera *speech community*, ciò al fine di ridurre gli sforzi e facilitare i confronti sul medesimo materiale. Due sono sostanzialmente gli aspetti che vincolano l'accessibilità di un database:
   a. il formato, che deve essere standardizzato ed aperto;
   b. principi etici, che concernono il suo utilizzo e la sua diffusione (maggiormente limitante nel caso di parlato emotivo naturale), a cui si aggiungono anche questioni di *copyright*.

È facile intuire come la generazione di un *database* che rispecchi le caratteristiche individuate dagli autori rappresenti un'impresa assai difficile: solo gli obiettivi e le finalità della specifica ricerca saranno in grado di stabilire e determinare le scelte e le caratteristiche individuate da Douglas-Cowie *et al.* (2003).

Prendendo invece in esame la ricognizione di quelle che Ververidis & Kotropoulos (2006) chiamano *emotional speech data collections*, o raccolte di parlato emotivo, emerge una verità che nel presente contesto assume una rilevante importanza.[21] Andando infatti alla ricerca di una risorsa di parlato emotivo mistilingue che risponda alle tematiche che qui si intendono affrontare, si nota l'assoluta assenza di simili risorse tranne qualche eccezione. Le uniche risorse mistilingui riscontrabili contengono stimoli in non più di due lingue: inglese e tedesco nel caso di Batliner *et al.* (2004), Scherer *et al.* (2002); inglese e sloveno nel caso di Ambrus (2000) e inglese e spagnolo in Gonzalez (1999). Il dato appena riportato non è assolutamente sorprendente se si considera poi che la sezione web dedicata da HUMAINE[22] alle risorse disponibili nell'ambito delle emozioni riporta, di fatto, 21 risorse per il solo parlato e 13 risorse contenenti materiale audio-visivo ed altro: l'elenco è decisamente meno aggiornato di quello di Ververidis & Kotropulos (2006).

Ciò che si vuole qui sottolineare non è tanto l'assenza di risorse linguistiche, a cui se ne sono nel frattempo aggiunte altre a cui probabilmente non si è riusciti ad avere accesso, quanto il fatto che le emozioni vocali rappresentino di per sé una tematica assai complessa: se nelle singole lingue si incontrano una serie di difficoltà e di problematiche, da quanto è

[21] È per le ragioni che seguiranno che, in un certo qual modo, la scelta di creare un corpus sperimentale di parlato emotivo mistilingue europeo risulta motivata, ma su questo aspetto si ritornerà in modo più approfondito nel prosieguo.

[22] Cfr. http://emotion-research.net/wiki/Databases.
HUMAINE (*Human-Machine Interaction Network on Emotion*) è un Network di eccellenza nato nel 2004 all'interno del 6° Programma Quadro EU con 33 partner provenienti da 14 paesi diversi.

stato fatto qui notare, tali questioni si riflettono e si complicano ulteriormente nelle ricerche di tipo cross-linguistico e cross-culturale.

Inutile sottolineare, infine, l'importanza delle risorse di parlato emotivo nello studio del parlato emotivo, dove l'uno non avrebbe ragione di essere senza l'altro.

4. OBIETTIVI

Sulla base delle questioni di ordine metodologico richiamate nei precedenti paragrafi, vengono di seguito riportati gli obiettivi di cui al presente lavoro che consistono nel:

- raccogliere, in via sperimentale,[23] un corpus mistilingue 'europeo' di parlato emotivo in lingua italiana, francese, inglese e tedesca;[24]
- motivare e illustrare le caratteristiche del corpus raccolto, con particolare riferimento al protocollo di elicitazione adottato, al tipo di materiale linguistico utilizzato ecc.;
- validare in termini percettivi con gruppi di ascoltatori il corpus raccolto;
- verificare e valutare l'influenza del protocollo di elicitazione in rapporto alla riconoscibilità e alla rappresentatività delle produzioni emotive raccolte;
- verificare e valutare l'influenza del protocollo di elicitazione in rapporto al soggetto registrato (naif vs. attore) e valutarne la resa in termini di materiale utile.

5. CREAZIONE DEL CORPUS EMOTIVO MISTILINGUE

Nei paragrafi che seguono verranno illustrate le scelte di tipo metodologico utilizzate per la raccolta e la creazione del corpus di parlato emotivo mistilingue per le quattro lingue europee.

5.1 La frase standard

Avendo in mente un confronto di produzioni vocali emotive in contesto cross-linguistico si è optato per una frase standard, anche definita *carrier sentence*[25], che consentisse una adeguata confrontabilità dei dati riducendo in *primis* la variabilità inter-parlatore nella produzione delle emozioni vocali. Al fine di garantire un certo livello di confrontabilità dei dati (sia a livello acustico, sia a livello percettivo) sono stati fissati *a-priori* dei *desiderata*, secondo cui la frase scelta doveva essere:

- semanticamente neutra se estrapolata da uno specifico contesto. Nonostante gli sforzi non è stato possibile soddisfare questo requisito per il semplice fatto che qualsiasi frase, generata e trasmessa durante un atto comunicativo, porta con sé una serie di informazioni che non è possibile escludere *a-priori*. Il semanticamente neutro è stato perciò ricondotto alla possibilità di trovare una frase che potesse essere inclusa in qualsiasi contesto emotivo in modo coerente con l'emozione intesa;
- coerente con le situazioni rappresentate per ciascuna emozione: la frase scelta doveva essere in linea con il costrutto della scena o della situazione presentata per ciascuna emozione affinché la stessa non creasse nel soggetto situazioni di dubbio o diffidenza;

[23] Per questo aspetto non è stato possibile rispondere appieno ai *desiderata* di Douglas-Cowie *et al.* (2003).

[24] Si veda a tal proposito quanto richiamato nel precedente § sulla disponibilità di *corpora* di parlato emotivo, con particolare riferimento a Ververidis & Kotropoulos (2006).

[25] O ancora *control cluster* come la definiscono Williams & Stevens (1972).

- di utilizzo comune in ciascuna lingua esaminata: la frase doveva rispettare quelle che sono rappresentate dalle consuetudini terminologiche e di costrutto di ciascuna lingua che la rendesse accettabile agli *encoder* prima ed ai *decoder* dopo;
- 'facile' da analizzare: si fa riferimento alla possibilità di individuare con un certo grado di facilità e sicurezza quelli che sono i confini frasali, nello specifico all'inizio della frase che sarebbe dovuto avvenire con un fono possibilmente sonoro. Tale requisito è motivato dalla possibilità di stabilire con esattezza, nella fase di segmentazione ed etichettatura, l'inizio della produzione all'interno del segnale registrato;[26]
- contenere una pausa: si è ritenuto in questo caso significativa la presenza e l'inserimento di una pausa all'interno della frase, al fine di valutare, nelle successive fasi, se nelle quattro lingue esaminate vi sia un diverso utilizzo o gestione dei tempi legati alle pause e ai silenzi.

Date queste premesse, e sulla base di precedenti esperienze di ricerca e scelte operate da altri studiosi,[27] con l'aiuto di docenti di madrelingua[28] si è passati all'individuazione di una frase standard nelle quattro lingue esaminate[29] giungendo alla seguente scelta:

-  Non è possibile. Non ci posso credere.
-  Oh là là. C'est incroyable.
-  It can't be. I cannot believe it.
-  Das ist nicht möglich. Ich kann es nicht glauben.

5.2 I testi emotigeni o scenari

Per l'elicitazione delle emozioni desiderate si è fatto riferimento al paradigma degli scenari. Gli scenari sono rappresentati sotto forma di testi narrativi, per la stesura dei quali ci siamo avvalsi di alcuni esempi presenti in letteratura. Nello specifico sono stati utilizzati, adattandoli in alcune parti, i testi emotigeni creati e riportati in Anolli & Ciceri (1992) e successivamente utilizzati anche in Anolli *et al.* (2008a, 2008b). L'adattamento dei testi si è

[26] Cosa che sarebbe sostanzialmente impossibile con foni di tipo occlusivo che sono caratterizzati per loro natura dalla presenza di una fase di occlusione, o di silenzio, non rilevabile in posizione di inizio di frase.

[27] Si fa qui riferimento a Walbott & Scherer (1986) che scelsero la frase *"Ich kann es nicht glauben."* e ad Anolli & Ciceri (1992) (ma anche Anolli *et al.* 2008a, 2008b) che scelsero la frase *"Non è possibile, non ora."*

[28] Tutti con esperienza nel campo dell'insegnamento e della traduzione e in servizio presso la Facoltà di Lettere e Filosofia dell'Università della Calabria.

[29] Scelta operata anche al fine di poter valutare il peso della competenza della lingua in un esperimento percettivo di tipo cross-linguistico sulle emozioni vocali.

reso necessario al fine di consentirne una più agevole traduzione nelle lingue oggetto di studio.[30]

Sono stati utilizzati complessivamente 6 scenari con riferimento alle emozioni definite da Ekman (1992) come *basic* (*happiness*, *anger*, *fear*, *sadness*, *disgust*, *surprise*). Poiché negli studi a cui si è fatto riferimento per gli scenari, l'emozione di sorpresa non era stata trattata, si è reso necessario procedere alla stesura *ex novo* del relativo testo seguendo le linee guida riportate nel paragrafo 3.2.4.

L'utilizzo degli scenari ha inoltre consentito di ridurre al minimo il ricorso ad etichette verbali di tipo emozionale (rabbia, tristezza ecc.) nelle quattro lingue, evitando così le problematiche a cui si è accennato al paragrafo 3.4.

5.3 Modalità di raccolta e caratteristiche del corpus

La raccolta del corpus è stata effettuata in ambiente insonorizzato con un microfono direzionale mod. *Sennheiser e835* direttamente in formato *.wav con l'ausilio di un registratore digitale mod. *Edirol R-09* a 44.1kHz 16-bit mono.[31] Il microfono, sorretto da un'asta microfonica, è stato posizionato ad una distanza di ca. 10 cm[32] e con un'angolatura di 45° rispetto alla fonte di emissione.

La procedura di raccolta è stata suddivisa in quattro fasi, una consecutiva all'altra.[33] Ciascuna delle quattro fasi risulta caratterizzata da una specifica modalità di acquisizione del materiale scaturita dalla somministrazione al soggetto di particolari istruzioni o informazioni.

Tutte le scelte, le strategie e le impostazioni metodologiche sono state inizialmente verificate e testate con l'aiuto di due attori italiani. Dopo una prima sommaria analisi del materiale acquisito, si sono rivelati necessari piccoli aggiustamenti nell'impostazione della procedura di raccolta. Di seguito le caratteristiche definitive di ciascuna delle modalità di raccolta del corpus in questione:

- Modalità A: al soggetto è stato chiesto di leggere i 6 brani (uno per ciascuna emozione studiata) in modo spontaneo e naturale senza ulteriori informazioni;
- Modalità B: dopo aver esplicitato al soggetto il motivo della registrazione e dopo un feedback da parte del soggetto riguardo l'emozione a cui si faceva riferimento in

[30] Particolare attenzione è stata data in questo caso all'utilizzo di termini di uso comune, evitando volutamente un linguaggio 'forbito' che avrebbe avuto ripercussioni sulla caratteristica di spontaneità in fase di produzione da parte dei soggetti.

[31] Le registrazioni sono state effettuate in parte nella camera insonorizzata (marca *Amplifon 2x2*) in dotazione al Laboratorio di Fonetica dell'Università della Calabria (http://www.linguistica.unical.it/labfon/Home.htm), in parte nella camera insonorizzata del KTH di Stoccolma durante un soggiorno del primo autore VG presso il *Centre for Speech Technology* (CTT) in qualità *guest researcher* (http://www.speech.kth.se/ctt/) e in parte presso lo studio di registrazione internazionale *Studio Colosseo* con sede in Roma (http://www.studiocolosseo.com/).

[32] Sebbene il microfono sia stato preventivamente posizionato alla stessa distanza per tutti i soggetti, mantenendo costante la distanza prefissata, è stato impossibile impedire ai soggetti di muoversi, causando un'oscillazione della distanza dal microfono di ±3cm.

[33] L'intera procedura ha richiesto tra i 40 e i 60 minuti a soggetto.

ciascun testo,[34] al soggetto è stato chiesto di rileggere il testo, sempre in modo spontaneo e naturale, rendendosi partecipe della situazione descritta (soprattutto nelle parti in cui era prevista la forma dialogica);

- Modalità C: al soggetto è stato chiesto di produrre la frase standard per ciascuna delle emozioni per almeno 4 volte.[35] Al soggetto è stato inoltre esplicitamente chiesto di fare riferimento alle emozioni descritte nei testi precedentemente letti: in questa fase è stato messo in atto quello che potrebbe essere definito uno pseudo - metodo Stanislavskij, sostituendo alla vita e all'esperienza personale richiesta dal metodo per l'autoinduzione dello stato emotivo, la descrizione contenuta nei brani presentati per ciascuna delle emozioni investigate.[36] A fine sessione è stata data facoltà al soggetto di ripetere qualora non fosse stato soddisfatto dalla propria *performance*;[37]
- Modalità Neutra: a conclusione delle tre fasi di raccolta delle produzioni verbali propriamente emotive, si è proceduto alla raccolta delle produzioni neutre per la stessa frase. Al soggetto è stato chiesto di leggere, per almeno quattro volte, ciascuna delle 5 frasi riportate su singoli fogli.[38] L'istruzione data al soggetto è stata quella di leggere le frasi nel modo più possibile neutro e senza alcuna caratterizzazione emotiva: l'esempio dato al soggetto è stato quello di leggere le frasi come se si trattasse di una lista di ingredienti per una ricetta culinaria.

In alcuni casi, a fine sessione di registrazione di ciascuna fase (prevalentemente per la modalità A e B), si è reso necessario far ripetere al soggetto qualcuno dei paragrafi contenente la frase standard, in quanto nella lettura erano stati inseriti nella frase standard suoni diversi da quelli previsti (in alcuni casi anche diverse parole).[39]

---

[34] Tale scelta è stata dettata dalla necessità di validare anche i brani adattati da Anolli & Ciceri (1992). È stato infatti in questa fase che è stato possibile riscontrare la presenza per il tedesco di due termini per quella che in italiano definiamo 'tristezza'. Nel caso specifico quella che era stata erroneamente indicata nel tedesco come *Traurigkeit* era invece *Trauer*.

[35] Al soggetto è stato chiesto di fare una breve pausa tra una produzione e l'altra al fine di evitare influenze di tipo stilistico causate dalla concatenazione delle frasi prodotte.

[36] È stato così possibile ridurre al minimo la variabilità legata alla situazione elicitante che in altri casi sarebbe stata inevitabilmente legata al vissuto personale dei singoli come prevede appunto il metodo Stanislavskij.

[37] Durante le prime sessioni di registrazione è stato possibile evidenziare come il soggetto si rendesse effettivamente conto delle produzioni prodotte in questa fase solo dopo aver esperito, attraverso la produzione verbale, le emozioni richieste, manifestando la volontà di ripetere alcune delle produzioni emotive realizzate e per le quali era stato espresso un certo grado di insoddisfazione.

[38] La frase di interesse è stata inserita all'interno dei 5 fogli e quindi delle 5 frasi. Anche in questo caso al soggetto è stato chiesto di fare una breve pausa tra una produzione e l'altra al fine di evitare influenze di tipo stilistico causate dalla concatenazione delle frasi prodotte.

[39] Per la caratteristica delle modalità di raccolta delle produzioni nella modalità A e B, sebbene si tratti di parlato letto, le stesse possono essere considerate come produzioni emotive spontanee o semi-spontanee dal momento che i soggetti coinvolti ignoravano quale fosse la frase oggetto di studio e quale fosse, di fatto, l'obiettivo specifico delle due modalità. Di contro, le produzioni raccolte nella modalità C possono essere considerate a

L'unica informazione fornita ai soggetti in fase di reclutamento riguardava esclusivamente l'oggetto di studio cross-linguistico delle emozioni vocali nell'ambito di una tesi di dottorato. Qualsiasi altra richiesta di informazioni in merito allo svolgimento dell'esperimento è stata declinata a tutela del *setting* sperimentale, spiegando che qualsiasi altra informazione aggiuntiva avrebbe vanificato il costrutto sperimentale della ricerca.

Per ciascuna lingua sono state acquisite registrazioni per le due categorie di soggetti, rispettivamente *naif* e *attori* professionisti, tutti di sesso maschile. La tabella n. 5 riporta un quadro relativo al numero e l'età media (tra parentesi la deviazione standard) dei soggetti *naif* e degli *attori* coinvolti in questa fase per l'acquisizione delle produzioni emotive in ciascuna lingua:

Lingua	Soggetti	Numero	Eta' media (SD)
Italiano (it)	attori	5	35,8 (5,2)
	naif	6	25,8 (1,6)
Francese (fr)	attori	3	43,7 (11,0)
	naif	6	29,5 (12,8)
Inglese (en)	attori	2	46,0 (8,5)
	naif	6	45,0 (10,9)
Tedesco (de)	attori	2	50,5 (2,1)
	naif	9	26,0 (3,9)

Tabella 5: Riepilogo informazioni relative al campione dei soggetti registrati

Prima e durante le sessioni di registrazione, tutte le istruzioni sono state fornite a ciascun soggetto nella propria lingua madre in forma sia scritta che orale.

La procedura di raccolta appena descritta è stata attuata per ciascun soggetto di ciascuna lingua per arrivare, a conclusione delle operazioni di etichettatura, ad un totale di 40 produzioni a soggetto così suddivise:

- Modalità A: 6 produzioni emotive (1 per emozione);
- Modalità B: 6 produzioni emotive (1 per emozione);
- Modalità C: 24 produzioni emotive (4 per ciascuna emozione);
- Modalità Neutra: 4 produzioni neutre.

Di seguito, alla fase di raccolta delle produzioni emotive per ciascun soggetto, si è provveduto ad isolare ed etichettare le singole produzioni contenute nel *continuum* delle registrazioni acquisite.

5.4 Etichettatura del corpus

L'etichettatura del corpus, intesa qui nell'accezione di isolamento mediante segmentazione della frase standard raccolta per ciascun soggetto nelle quattro modalità, è stata effettuata manualmente con ascolto in cuffia e con l'ausilio di un software *open source* denominato Praat.[40] In questa fase particolare attenzione è stata prestata

pieno titolo come produzioni emotive posate o simulate, in quanto in tale modalità l'attenzione del soggetto è stata fatta convergere esclusivamente sulla frase proposta.

[40] Cfr. Boersma, P. & Weenink, D. (2009), *Praat: doing phonetics by computer* [Computer program], retrieved from http://www.praat.org/. Il programma, oltre a consentire di visualizzare in contemporanea oscillogramma e sonogramma del segnale analizzato

all'individuazione dei punti di attacco e fine frase, avendo cura di operare la segmentazione della porzione o, più brutalmente, del taglio del segnale sonoro, in corrispondenza dello *zero-crossing* (v. fig. 2).

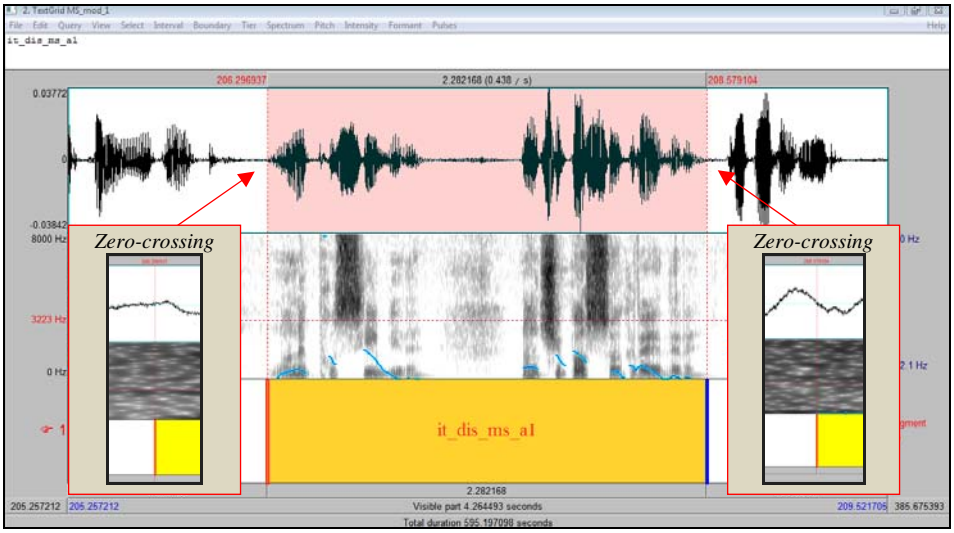


Figura 2: Esempio di schermata relativa alla segmentazione e all'etichettatura delle produzioni

L'etichettatura della porzione di segnale di interesse fornisce il nome del file assegnato alla porzione nella fase di esportazione della stessa in un singolo file. L'etichettatura adottata in fase di segmentazione codifica tutte le informazioni disponibili per quella data porzione, informazioni standardizzate secondo uno schema a sei campi in termini di [lingua]_[emozione]_[categoria]_[iniziali]_[(modalità)occorrenza].wav, dove:

1. [lingua] = campo di due lettere che corrisponde alla lingua codificata:
   - it (italiano);
   - fr (francese);
   - en (inglese);
   - de (tedesco);
2. [emozione] = campo di tre lettere che corrisponde all'emozione codificata:
   - ang (collera calda);
   - dis (disgusto);
   - fea (paura);
   - joy (gioia);
   - neu (neutro);
   - sad (tristezza);
   - sur (sorpresa);

---

rendendo più agevole e precisa la segmentazione e l'etichettatura, permette anche di non intaccare e/o modificare il file originale, riportando tutte le operazioni effettuate sull'originale in un file di testo denominato *Textgrid*.

3. [categoria] = campo di una lettera che corrisponde alla categoria di appartenenza del soggetto utilizzato per l'*encoding* dell'emozione :
    - a (attore);
    - n (naif);
4. [iniziali] = campo di due lettere corrispondente alle iniziali del soggetto registrato;
5. (modalità) = campo di una lettera che corrisponde alla modalità di raccolta della produzione emotiva codificata:
    - a (lettura del brano in Modalità A);
    - b (lettura del brano in Modalità B);
    - c (lettura della sola frase in Modalità C);
    - campo non contemplato nel caso di produzione neutra in quanto raccolto in un'unica e sola modalità (Modalità Neutra);
6. [occorrenza] = campo di una cifra relativa all'occorrenza della frase prodotta.

A titolo esemplificativo vengono di seguito forniti alcuni esempi relativi alla predetta etichettatura dei *files*: it_neu_n_gb_1.wav; fr_ang_a_hd_c3.wav e così via.

A conclusione della fase di segmentazione ed etichettatura dell'intero corpus sono risultate complessivamente 1560 produzioni come di seguito meglio dettagliato (v. Tab. 6):

	Soggetti naif	Attori	Totale	N° produzioni
Italiano (it)	6	5	11	440
Francese (fr)	6	3	9	360
Inglese (en)	6	2	8	320
Tedesco (de)	9	2	11	440
Totale	**27**	**12**	**39**	**1560**

Tabella 6: Riepilogo delle registrazioni complessivamente effettuate

6. PROCEDURA DI VALIDAZIONE DEL CORPUS

Come sottolinea Magno Caldognetto (2002: 212), le registrazioni acquisite devono necessariamente essere validate attraverso dei test percettivi e di adeguatezza al fine di "selezionare le produzioni riconosciute esattamente e con il punteggio più elevato da parte degli ascoltatori […]".

Ciò si rende necessario non solo per assicurare la significatività di tutte quelle operazioni legate al prosieguo della ricerca e della sperimentazione intrapresa (misurazioni acustiche, prove percettive ecc.), ma anche per assicurare e garantire un certo grado di significatività e confrontabilità dei risultati con quelli conseguiti in altre ricerche. D'altro canto, tale operazione si rende assolutamente necessaria nel caso si presentino produzioni emotive di una lingua a membri appartenenti ad un'altra lingua o cultura.

Anche nel caso della validazione allo stato attuale non esistono, per quanto sia noto agli autori, procedure standardizzate. Per questa ragione, per la validazione delle registrazioni acquisite per ciascuna lingua, sono stati somministrati due test di adeguatezza:

- uno relativo all'identificazione dell'emozione espressa nello stimolo presentato (nel prosieguo identificato per comodità con la sigla 'T1') con una scelta da effettuare sulla base di un set prefissato di etichette emozionali (tristezza, collera, paura, neutro, gioia, sorpresa e disgusto) con in aggiunta la possibilità di esprimere il grado

di certezza della risposta data (incerto, certo)[41]. Scopo di questo test è stato quello di escludere tutte quelle produzioni che potevano suscitare confusione e che potevano essere considerate emotivamente ambigue;

- uno relativo alla rappresentatività dello stimolo presentato (nel prosieguo identificato con la sigla 'T2') contestualmente all'indicazione dell'emozione intesa in fase di *encoding* (rappresentatività espressa su una scala di giudizio di tipo pari a quattro gradi).[42]

La scelta di somministrare un secondo test di validazione, come quello sulla rappresentatività di un determinato stimolo in funzione dell'emozione intesa, è riconducibile a due motivazioni:

- la prima è sostanzialmente dettata dall'aver in mente una ulteriore selezione degli stimoli da utilizzare in un secondo esperimento percettivo di tipo cross-linguistico (e per il quale tutto il lavoro qui presentato si è reso necessario): a parità di identificazione corretta di una o più produzioni per una stessa emozione espressa da un soggetto, sarebbe stato così possibile scegliere quella giudicata più rappresentativa da parte dei giudici ascoltatori;
- la seconda è dovuta alla verifica di una ipotesi, ovvero che uno stimolo non correttamente identificato difficilmente avrebbe avuto una valutazione alta in termini di rappresentatività.

Per una validazione omogenea dell'intero *corpus* e per evitare di somministrare troppi stimoli ad un unico soggetto,[43] e poiché per ciascuna lingua sono stati utilizzati gli stessi soggetti che hanno partecipato alla fase di acquisizione, a ciascun soggetto sono stati somministrati i due test includendo le registrazioni prodotte dal soggetto ascoltatore medesimo e quelle di altri due soggetti per la stessa lingua. I test contenenti gli stimoli da far ascoltare ai soggetti sono stati generati sulla base di una matrice come quella presentata in Tabella 7.

DECODER		**A[1]**	**B[1]**	**C[1]**	**D[1]**	**E[1]**
ENCODER	**A**					
	B					
	C					
	D					
	E					

Tabella 7: Modalità di generazione dei set di ascolto utilizzati per i due test di validazione (ordine di lettura dall'alto verso il basso)

Si supponga, per ciascuna lingua, la presenza di un set di voci A, B, C, D, E come riportato in Tabella 7, dove la colonna di sinistra si riferisce alle produzioni verbali emotive, mentre la riga in alto si riferisce all'ascoltatore (che nel presente caso, come nei test successivamente condotti, coincidono con gli *encoder*). Sulla base della matrice

[41] Trattandosi di un test a risposta forzata è stata data agli ascoltatori la possibilità di esprimere il grado di confidenza (certezza/incertezza) sulla risposta data al fine di evitare che gli stessi identificassero con l'opzione 'neutro' la condizione di incertezza o di dubbio.

[42] È stata qui adottata una scala di giudizio *pari* per obbligare il soggetto a prendere una posizione chiara ed univoca evitando situazioni di dubbio o di neutralità.

[43] Si consideri che per ciascun soggetto sono stati complessivamente acquisiti 40 stimoli.

riportata in Tabella 7, le produzioni del soggetto A, oltre ad essere ascoltate da se stesso (A^1) vengono ascoltate dai soggetti B^1 ed E^1; le produzioni del soggetto B, oltre ad essere ascoltate da se stesso (B^1) vengono ascoltate dai soggetti A^1 e C^1; le produzioni del soggetto C, oltre ad essere ascoltate da se stesso (C^1) vengono ascoltate dai soggetti B^1 e D^1; le produzioni del soggetto D, oltre ad essere ascoltate da se stesso (D^1) vengono ascoltate dai soggetti C^1 ed E^1; infine, le produzioni del soggetto E, oltre ad essere ascoltate da se stesso (E^1) vengono ascoltate dai soggetti D^1 e A^1 chiudendo in questo modo la batteria dei test. In questo modo al soggetto A^1 verranno somministrate le produzioni sue (quindi A) nonché quelle di B ed E e così via. Secondo tale costrutto, ciascuna produzione di ciascun soggetto viene valutata complessivamente da tre soggetti.

I due test sono stati quindi implementati con la funzione 'ExperimentMFC' del software Praat e somministrati in cuffia: gli stimoli sono stati opportunamente randomizzati con una apposita *routine* all'interno dello stesso programma.[44] In entrambi i casi il test era preceduto da una schermata con poche e semplici istruzioni sul compito da svolgere. Anche in questo caso il tutto è stato svolto nella lingua madre dei soggetti ascoltatori (v. Fig. 3).

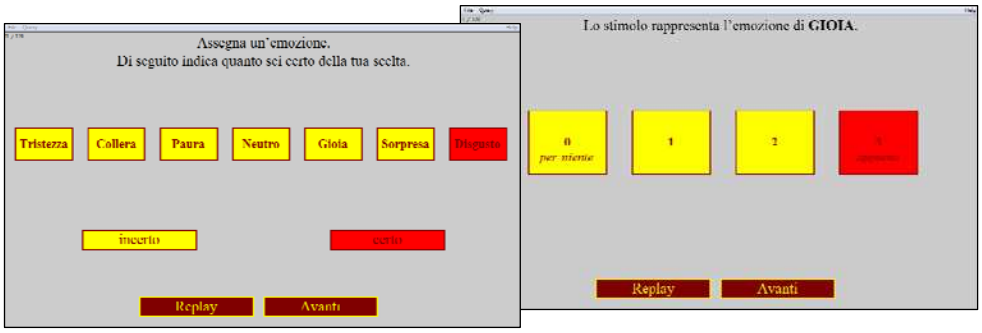


Figura 3: Schermate relative rispettivamente al test di validazione T1 sull'identificazione dell'emozione presentata e T2 sulla rappresentatività dell'emozione presentata

7. RISULTATI

Vengono di seguito riportati i risultati relativi ai due test di validazione adottati per verificare la bontà e l'adeguatezza delle registrazioni di parlato emotivo raccolte per ciascuna delle quattro lingue secondo le modalità sopra riportate, registrazioni che costituiranno il corpus di parlato emotivo mistilingue qui presentato.

Prima di intraprendere qualsiasi ulteriore analisi ci si è chiesti quanto i risultati dei due test di validazione, rispettivamente T1 e T2, fossero tra loro correlati. Poiché era stato ipotizzato che uno stimolo non correttamente identificato nel T1 difficilmente avrebbe avuto una valutazione alta in termini di rappresentatività nel T2, per verificare la validità di tale ipotesi si è proceduto al calcolo di un coefficiente di correlazione sui valori medi delle risposte fornite per ciascuno stimolo nei due test di validazione: per tutte e quattro le lingue è stata rilevata una correlazione superiore a 0.5 con un livello di significatività $p = 0.01$

[44] Per evitare il calo di attenzione da parte dei giudici ascoltatori, in entrambi i test, è stata prevista una pausa ogni 30 stimoli presentati.

supportando quella che era l'ipotesi di partenza: rispettivamente r = 0.672 per l'italiano, r = 0.706 per il francese, r = 0.532 per l'inglese e r = 0.705 per il tedesco.

7.1 Test di validazione relativo all'identificazione delle emozioni (T1)

Il primo dato ad emergere dai test di validazione relativi all'identificazione delle emozioni presentate ai giudici ascoltatori è sicuramente legato alla differenza dei valori di corretta identificazione delle emozioni espresse in base alla modalità di raccolta delle produzioni verbali emotive. Infatti, come si evince dalle matrici di confusione generate sulla base del test di identificazione T1, per ciascuna delle quattro lingue esaminate e per ciascuna delle modalità di raccolta delle produzioni (vedi Tab. 8-11), emerge sostanzialmente come vi sia, da parte dei giudici ascoltatori, un certo grado di abilità nell'identificare produzioni emotive nella propria lingua con percentuali nettamente al di sopra della pura casualità.[45] Indifferentemente dalle modalità, si registrano valori medi di corretta identificazione del 56% per l'italiano, del 51% per il francese, del 44% per l'inglese e del 49% per il tedesco. Si tratta di percentuali indubbiamente alte se si considera il fatto che si tratti, in questo caso, di materiale presentato ai giudici ascoltatori in forma 'grezza'.

ITALIANO

Valori complessivi

	sad	ang	fea	neu	joy	sur	dis
sad	**74%**	2%	4%	14%	1%	2%	5%
ang	2%	**63%**	3%	14%	1%	7%	13%
fea	12%	2%	**44%**	15%	4%	13%	11%
neu	6%	1%	0%	**90%**	2%	0%	2%
joy	4%	3%	8%	8%	**40%**	35%	3%
sur	9%	4%	6%	16%	5%	**40%**	20%
dis	6%	13%	5%	13%	3%	21%	**39%**

Modalità di raccolta A

	sad	ang	fea	neu	joy	sur	dis
sad	**52%**	3%	6%	27%	0%	3%	9%
ang	6%	**15%**	6%	42%	0%	9%	21%
fea	3%	0%	**15%**	42%	6%	21%	12%
neu	6%	1%	0%	**90%**	2%	0%	2%
joy	9%	0%	6%	24%	**21%**	33%	6%
sur	6%	0%	9%	30%	0%	**36%**	18%
dis	3%	15%	3%	33%	0%	33%	**12%**

Modalità di raccolta B

	sad	ang	fea	neu	joy	sur	dis
sad	**76%**	0%	9%	15%	0%	0%	0%
ang	3%	**45%**	3%	18%	3%	9%	18%
fea	6%	3%	**42%**	15%	3%	12%	18%
neu	6%	1%	0%	**90%**	2%	0%	2%
joy	3%	3%	9%	15%	**42%**	24%	3%
sur	15%	6%	6%	18%	0%	**27%**	27%
dis	0%	15%	6%	18%	0%	18%	**42%**

Modalità di raccolta C

	sad	ang	fea	neu	joy	sur	dis
sad	**80%**	2%	2%	10%	1%	2%	5%
ang	0%	**79%**	2%	5%	0%	5%	9%
fea	16%	2%	**52%**	8%	4%	11%	8%
neu	6%	1%	0%	**90%**	2%	0%	2%
joy	2%	3%	8%	2%	**45%**	38%	2%
sur	8%	5%	5%	12%	8%	**45%**	18%
dis	6%	9%	0%	9%	6%	18%	**52%**

Tabella 8: Matrici di confusione relative al test di identificazione T1 per il sottocorpus mistilingue italiano[46]

[45] Pari a ca. il 14,3% in un test con sette opzioni.

[46] I valori di corretta identificazione sono riportati in diagonale per ciascuna delle modalità di raccolta con un ordine di lettura da sinistra a destra. Il totale per ciascuna riga può non corrispondere al 100% in quanto vengono riportati in tabella valori approssimati. Lo stesso vale per le tabelle che seguono.

FRANCESE

Valori complessivi

	sad	ang	fea	neu	joy	sur	dis
sad	**67%**	3%	6%	14%	2%	6%	3%
ang	1%	**56%**	5%	11%	5%	17%	7%
fea	22%	2%	**38%**	11%	3%	22%	1%
neu	6%	0%	3%	**82%**	0%	4%	6%
joy	4%	6%	5%	10%	**38%**	34%	3%
sur	4%	6%	4%	11%	12%	**52%**	12%
dis	14%	20%	6%	12%	2%	22%	**25%**

Modalità di raccolta A

	sad	ang	fea	neu	joy	sur	dis
sad	**44%**	7%	7%	33%	0%	4%	4%
ang	4%	**22%**	11%	30%	0%	15%	19%
fea	0%	0%	**44%**	30%	0%	26%	0%
neu	6%	0%	3%	**82%**	0%	4%	6%
joy	7%	4%	19%	30%	**11%**	30%	0%
sur	11%	7%	4%	26%	4%	**37%**	11%
dis	7%	26%	4%	30%	0%	22%	**11%**

Modalità di raccolta B

	sad	ang	fea	neu	joy	sur	dis
sad	**56%**	0%	7%	26%	0%	7%	4%
ang	0%	**56%**	4%	15%	7%	15%	4%
fea	30%	4%	**33%**	15%	11%	7%	0%
neu	6%	0%	3%	**82%**	0%	4%	6%
joy	4%	15%	0%	0%	**44%**	30%	7%
sur	7%	4%	4%	11%	11%	**44%**	19%
dis	15%	19%	7%	11%	4%	22%	**22%**

Modalità di raccolta C

	sad	ang	fea	neu	joy	sur	dis
sad	**75%**	3%	5%	6%	4%	6%	3%
ang	0%	**64%**	4%	5%	6%	18%	5%
fea	26%	3%	**38%**	6%	2%	24%	2%
neu	6%	0%	3%	**82%**	0%	4%	6%
joy	3%	5%	3%	7%	**44%**	36%	3%
sur	2%	6%	4%	7%	14%	**57%**	10%
dis	16%	19%	6%	7%	3%	21%	**29%**

Tabella 9: Matrici di confusione relative al test di identificazione T1 per il sottocorpus mistilingue francese.

INGLESE

Valori complessivi

	sad	ang	fea	neu	joy	sur	dis
sad	**60%**	1%	15%	13%	0%	3%	8%
ang	3%	**49%**	8%	6%	4%	15%	16%
fea	28%	1%	**38%**	13%	4%	4%	11%
neu	31%	3%	2%	**60%**	1%	0%	2%
joy	2%	6%	8%	14%	**26%**	37%	8%
sur	8%	4%	9%	7%	14%	**49%**	9%
dis	13%	10%	19%	8%	10%	17%	**23%**

Modalità di raccolta A

	sad	ang	fea	neu	joy	sur	dis
sad	**33%**	4%	25%	21%	0%	0%	17%
ang	13%	**25%**	8%	17%	4%	8%	25%
fea	21%	4%	**29%**	21%	4%	4%	17%
neu	31%	3%	2%	**60%**	1%	0%	2%
joy	4%	8%	8%	33%	**4%**	42%	0%
sur	8%	13%	8%	21%	8%	**29%**	13%
dis	13%	21%	17%	13%	8%	25%	**4%**

Modalità di raccolta B

	sad	ang	fea	neu	joy	sur	dis
sad	**67%**	0%	13%	13%	0%	4%	4%
ang	0%	**63%**	4%	8%	8%	13%	4%
fea	4%	0%	**50%**	21%	8%	4%	13%
neu	31%	3%	2%	**60%**	1%	0%	2%
joy	0%	0%	13%	8%	**17%**	54%	8%
sur	4%	8%	8%	4%	0%	**67%**	8%
dis	4%	8%	29%	4%	17%	17%	**21%**

Modalità di raccolta C

	sad	ang	fea	neu	joy	sur	dis
sad	**66%**	0%	14%	10%	0%	3%	7%
ang	2%	**51%**	8%	2%	3%	17%	17%
fea	35%	1%	**38%**	9%	3%	4%	9%
neu	31%	3%	2%	**60%**	1%	0%	2%
joy	2%	6%	7%	10%	**33%**	31%	9%
sur	9%	1%	9%	4%	19%	**49%**	8%
dis	15%	7%	18%	8%	9%	15%	**28%**

Tabella 10: Matrici di confusione relative al test di identificazione T1 per il sottocorpus mistilingue inglese

TEDESCO

Valori complessivi

	sad	ang	fea	neu	joy	sur	dis
sad	**66%**	1%	10%	16%	1%	4%	4%
ang	5%	**59%**	5%	10%	3%	10%	9%
fea	17%	6%	**40%**	12%	2%	16%	7%
neu	11%	2%	2%	**83%**	1%	2%	1%
joy	3%	14%	6%	9%	**35%**	27%	7%
sur	14%	10%	8%	11%	7%	**40%**	10%
dis	15%	15%	13%	11%	8%	20%	**18%**

Modalità di raccolta A

	sad	ang	fea	neu	joy	sur	dis
sad	**58%**	0%	3%	24%	0%	3%	12%
ang	3%	**42%**	3%	24%	3%	3%	21%
fea	9%	6%	**27%**	3%	3%	30%	21%
neu	11%	2%	2%	**83%**	1%	2%	1%
joy	3%	9%	15%	21%	**24%**	21%	6%
sur	15%	12%	18%	15%	0%	**24%**	15%
dis	18%	9%	18%	21%	0%	15%	**18%**

Modalità di raccolta B

	sad	ang	fea	neu	joy	sur	dis
sad	**67%**	0%	12%	9%	0%	6%	6%
ang	15%	**61%**	3%	6%	0%	12%	3%
fea	18%	6%	**52%**	12%	0%	9%	3%
neu	11%	2%	2%	**83%**	1%	2%	1%
joy	6%	15%	6%	9%	**21%**	30%	12%
sur	21%	9%	6%	6%	3%	**45%**	9%
dis	21%	21%	9%	6%	12%	18%	**12%**

Modalità di raccolta C

	sad	ang	fea	neu	joy	sur	dis
sad	**67%**	1%	11%	15%	1%	4%	1%
ang	3%	**63%**	5%	8%	3%	11%	7%
fea	19%	5%	**41%**	14%	2%	14%	5%
neu	11%	2%	2%	**83%**	1%	2%	1%
joy	2%	15%	3%	6%	**41%**	27%	6%
sur	11%	10%	6%	11%	10%	**43%**	9%
dis	6%	15%	15%	15%	9%	12%	**27%**

Tabella 11: Matrici di confusione relative al test di identificazione T1 per il sottocorpus mistilingue tedesco

Esaminando i valori complessivi di corretta identificazione, si rileva, citando Scherer, Johnstone & Klasmeyer (2003: 444), come le emozioni "Sadness and anger are generally best recognized vocally, followed by fear" con valori medi di corretta identificazione tra i più alti in tutte e quattro le lingue dopo le produzioni neutre, seguiti da gioia e disgusto (per le quali si rilevano, invece, le percentuali più basse). Ciò confermerebbe, quindi, quanto affermato da Johnstone & Scherer (2000), secondo i quali, il contesto dell'evoluzione nei termini proposti da Darwin avrebbe selezionato e favorito per alcune emozioni maggiori caratterizzazioni nella sfera visiva, mentre per altre nella sfera acustica: per la rabbia e la paura sarebbe stata maggiormente sviluppata l'espressione vocale, perché gli antenati dell'uomo potessero avvertirsi e minacciarsi in modo esplicito anche a lunghe distanze; per emozioni quali disgusto e gioia sarebbero, invece, stati selezionati e favoriti sviluppi biologici relativi alla mimica facciale per far sì che, fra membri appartenenti alla stessa comunità, ci si potesse capire al volo.

Al di là delle matrici di confusione, per quel che concerne il presente test di validazione sono state considerate come utili tutte quelle produzioni (*item*) in cui almeno 2 dei 3 giudici ascoltatori hanno correttamente identificato l'emozione intesa in fase di *encoding* secondo le procedure di raccolta di cui al § 5. Sulla base di tale filtro (2 su 3) sono state complessivamente ritenute utili rispettivamente il 55,2% delle produzioni per l'italiano, il 48,1% per il francese, il 43,8% per l'inglese e il 45% per il tedesco, la cui composizione in termini di produzioni per modalità di raccolta, risulta nelle seguenti proporzioni (v. Fig. 4):

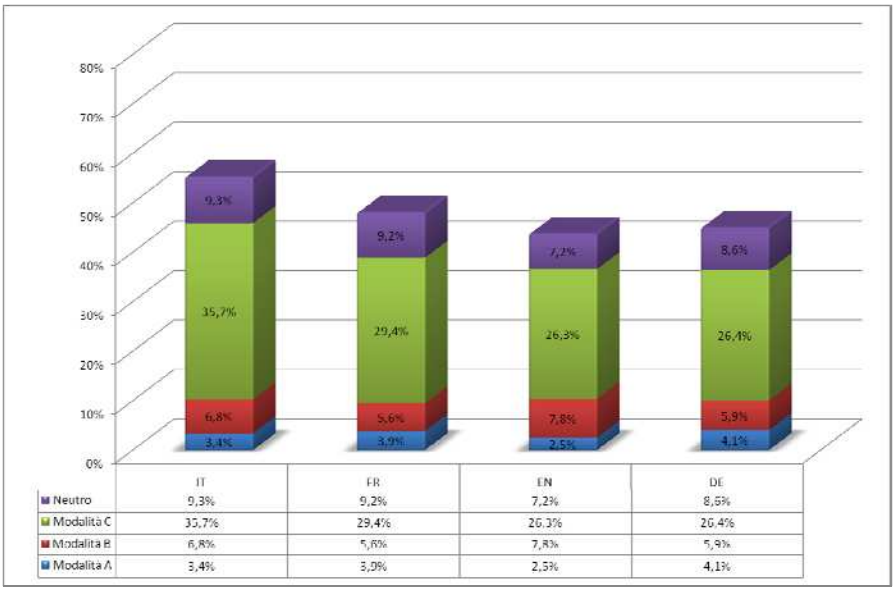


Figura 4: Consistenza del corpus per modalità di raccolta e per lingua in base al test di validazione T1

Se si esaminano, invece, i valori relativi a ciascuna modalità di raccolta (v. Fig. 5), emerge come i giudici ascoltatori di ciascuna lingua abbiano identificato con maggiore facilità le produzioni neutre seguite dalle produzioni raccolte in modalità isolata (modalità C), per finire con le più difficili da identificare nella modalità di raccolta A, confermando come la modalità A produca, date le sue caratteristiche, uno scarso coinvolgimento dell'*encoder* in quello che è rappresentato dallo scenario fornito.

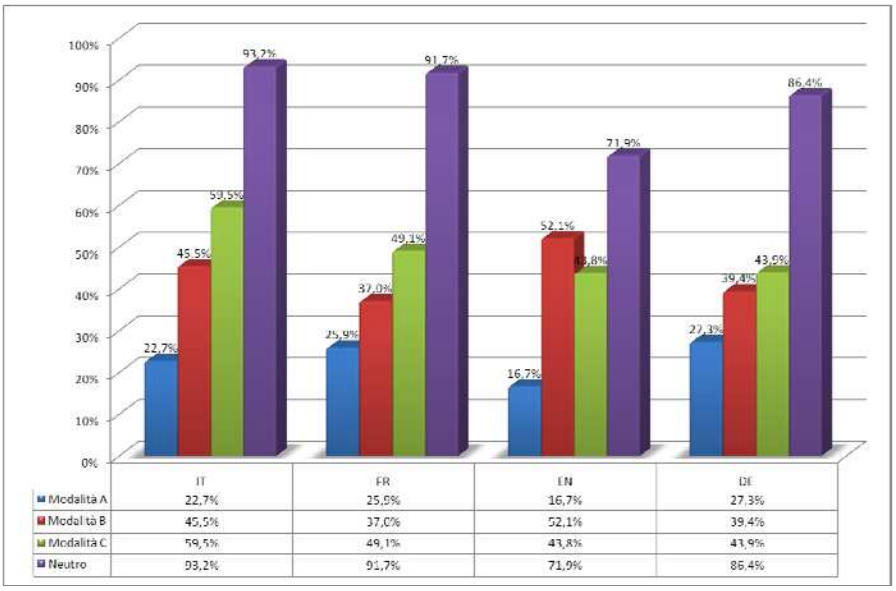


Figura 5: Materiale ritenuto utile per ciascuna lingua e per ciascuna modalità in base al test di validazione T1

Ad ogni modo, quali che siano le modalità di raccolta, si evince chiaramente come la validazione nei termini proposti da Magno Caldognetto (2002) sia non solo auspicabile, ma assolutamente necessaria al fine di stabilire con un buon grado di certezza ciò che si sta analizzando.

7.2 Test di validazione relativo alla rappresentatività delle emozioni (T2)

Per il test di validazione relativo alla rappresentatività delle emozioni presentate in riferimento all'emozione intesa in fase di *encoding*, prima di passare all'analisi dei risultati scaturiti da questa fase, si è proceduto alla verifica dell'attendibilità dei giudici ascoltatori in merito alla coerenza delle risposte fornite per uno stesso *item* (in questo caso rappresentati dai singoli *files*).

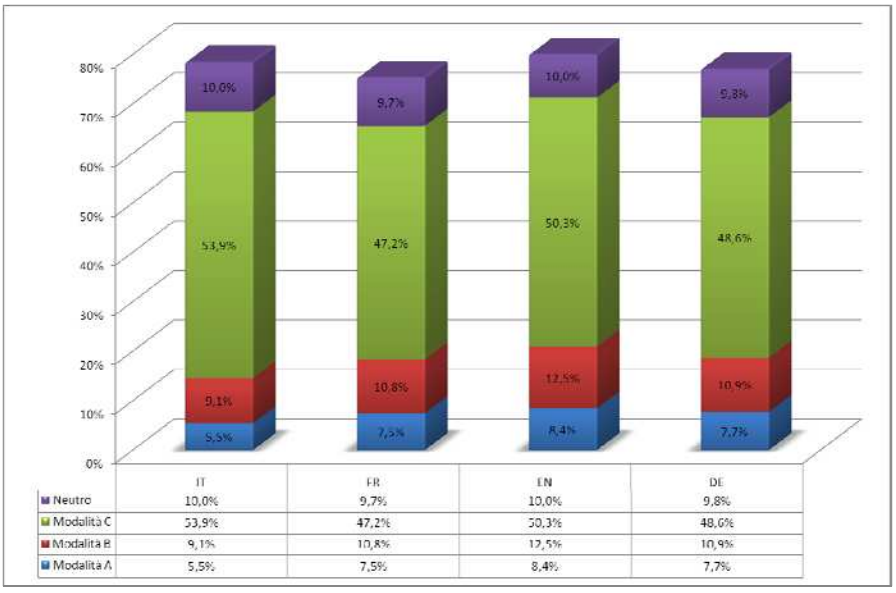
Poiché il test è stato costruito di modo che ciascun set di 40 produzioni di ciascun *encoder* venisse ascoltato da tre giudici ascoltatori (compreso l'*encoder* delle produzioni presentate), per ciascuna lingua e per ciascuno di tali set si proceduto al calcolo del coefficiente *alfa di Cronbach* con l'ausilio del software statistico SPSS.[47]

I valori medi rilevati per il coefficiente *alfa di Cronbach* sono: 0.672 per l'italiano, 0.640 per il francese, 0.558 per l'inglese e 0.636 per il tedesco. Va, tuttavia, fatto notare che il valore del coefficiente *alfa di Cronbach* cresce all'aumentare del numero degli *item*

[47] Si veda a tal proposito la Tabella 7, dove il calcolo del coefficiente *alfa di Cronbach* è stato calcolato in modo trasversale a quello di generazione dei set di ascolto, ovvero da sinistra verso destra: ad es. con riferimento al set delle produzioni dell'*encoder* B in Tabella 7, il coefficiente *alfa di Cronbach* è stato calcolato sulle risposte fornite rispettivamente da A^1, B^1 e C^1 procedendo in modo analogo per tutte le altre.

considerati: il numero ridotto di *item* (40) su cui il coefficiente è stato calcolato potrebbe in tal caso aver influito sulla restituzione di valori relativamente bassi. Date le finalità qui perseguite e dal momento che lo scopo primario del T2 è quello di selezionare il materiale considerato più rappresentativo per le emozioni intese, i valori riscontrati, sebbene bassi, sono comunque da ritenersi sufficienti.[48]

Assumendo come soglia di significatività un valore di rappresentatività pari o superiore a 2.3 per ciascun *item* dei test (su una scala di giudizio a 4 passi con un massimo di 4) è stato complessivamente ritenuto utile il 78,4% delle produzioni raccolte per il sotto corpus italiano, il 75,3% per quello francese, l'81,3% per quello inglese e il 77% per quello tedesco. La consistenza interna dei sotto *corpora* che compongono il corpus mistilingue qui raccolto in base alle modalità di raccolta più sopra esplicitate, risulta altrettanto simile nelle quattro lingue prese in considerazione (v. Fig. 6).



	IT	FR	EN	DE
Neutro	10,0%	9,7%	10,0%	9,8%
Modalità C	53,9%	47,2%	50,3%	48,6%
Modalità B	9,1%	10,8%	12,5%	10,9%
Modalità A	5,5%	7,5%	8,4%	7,7%

Figura 6: Consistenza del corpus per modalità di raccolta e lingua in base al test di validazione T2

Sebbene la parte più cospicua e presente all'interno dei vari sotto *corpora* (che compongono il corpus mistilingue) sia rappresentata in termini assoluti dalla modalità di raccolta C (modalità per la quale è stato possibile raccogliere più produzioni), è interessante rilevare come, in termini relativi, le produzioni appartenenti alla modalità Neutra e alla modalità di raccolta C siano le produzioni considerate dai giudici ascoltatori tra le più rappresentative come evidenzia il grafico riportato in Fig. 7.

[48] Come riportano Barbaranelli & D'Olimpio (2007: 241) non esistono regole statistiche per l'interpretazione del coefficiente di attendibilità, "ma si segue una regola pratica secondo la quale valori uguali almeno a .90 vengono considerati ottimi, valori compresi tra .80 e .90 molto buoni, valori compresi tra .70 e .80 buoni, valori compresi tra .60 e .70 sufficienti, valori inferiori a .60 inadeguati".

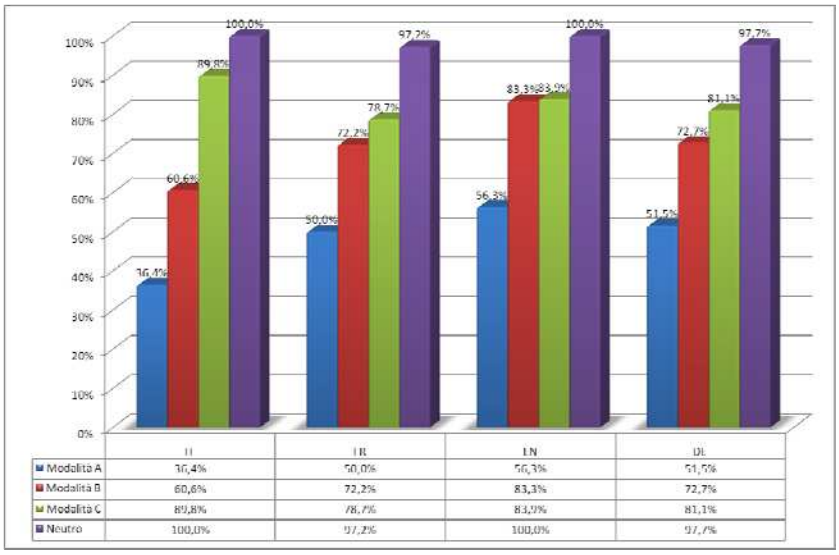


Figura 7: Materiale utile per ciascuna lingua e per ciascuna modalità di raccolta in base al test di validazione T2

7.3 Quale modalità di raccolta

Un altro obiettivo della presente proposta consisteva nel verificare e valutare l'influenza del protocollo di elicitazione in rapporto alla riconoscibilità e alla rappresentatività delle produzioni emotive raccolte.

Dalle analisi sopra riportate emerge chiaramente come la modalità di raccolta C (che può essere considerata come raccolta di produzioni emotive posate) sia di gran lunga la più efficace se si considera la quantità di materiale utile scaturita dalle procedure di validazione adottate in questa sede.

Sebbene le modalità di raccolta che sono state qui etichettate come modalità di raccolta A e B, restituiscano anch'esse produzioni riconoscibili e sufficientemente rappresentative delle emozioni intese, vi sono sostanzialmente due grossi limiti nel loro utilizzo: se, da un lato, il dispendio in termini di sforzi e di tempo nella raccolta del materiale sonoro diventa inevitabilmente estenuante, dall'altro si rischia di non disporre, a fine raccolta e a fine validazione, del materiale necessario allo svolgimento della ricerca avviata. Come si rileva dai dati riportati in Figura 5 per le due modalità A e B rispetto al test di validazione T1, e se si rapportano le percentuali al numero di produzioni emotive effettivamente utili (rispettivamente 15 su 66 e 30 su 66 per l'italiano, 14 su 54 e 20 su 54 per il francese, 8 su 48 e 25 su 48 per l'inglese e 18 su 66 e 26 su 66 per il tedesco), è facile intuire come non sia possibile rappresentare, per ciascun soggetto, un set completo composto di 6 produzioni emotive (ad es. per l'italiano per la modalità A sono state valutate utili solo 15 produzioni emotive su un totale di 66 relative agli 11 soggetti registrati).[49]

[49] I dati qui riportati sono relativi alle produzioni utili ricavate sulla base di soglie di significatività individuate nei precedenti paragrafi.

Va parimenti sottolineato come assai probabilmente la sequenza di acquisizione delle registrazioni, secondo le modalità a cui si è più volte accennato, caratterizzate da un livello di informazione e coinvolgimento del soggetto via via crescente, abbia aiutato i soggetti a focalizzarsi meglio sulle emozioni demandate nella modalità C. Da questo punto di vista la procedura adottata potrebbe addirittura essere considerata propedeutica per far esperire al soggetto le emozioni intese in ciascuno degli scenari presentati.

7.4 Soggetti naif o attori

Un altro aspetto qui indagato mirava a verificare e valutare l'influenza del protocollo di elicitazione in rapporto al soggetto registrato (naif vs. attore) valutandone la resa in termini di materiale utile. Senza ombra di dubbio l'utilizzo di attori produce sempre e comunque risultati migliori in termini di produzioni correttamente identificate e in termini di produzioni rappresentative.

Ne è un esempio il grafico riportato in Fig. 8 relativo ai due test in cui si rileva come, a parità di condizioni (procedure e protocollo di raccolta), l'ago della bilancia penda complessivamente a favore degli attori, nonostante qualche trascurabile differenza tra le quattro lingue.

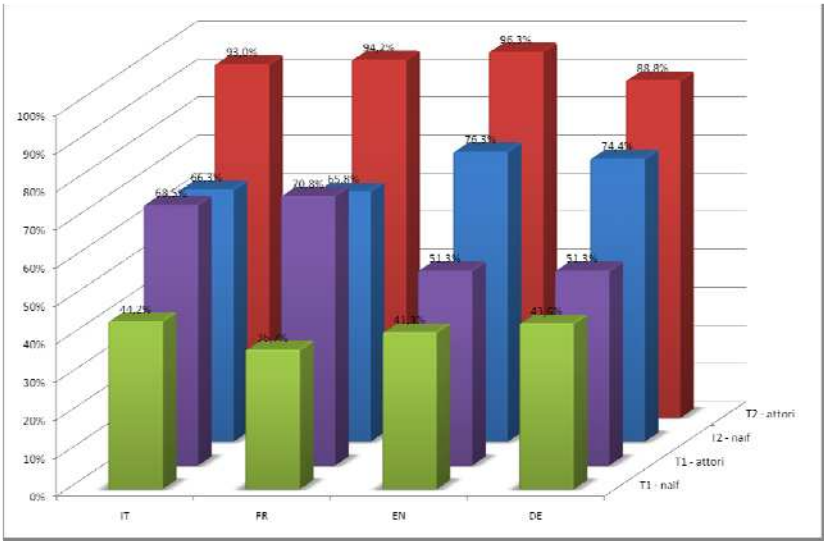


Figura 8: Produzioni ritenute utili per i due test di validazione con riferimento alle due categorie di *encoder* (*naif* vs. *attori*)

Nel caso di ricerche di tipo sperimentale sul parlato emotivo, l'utilizzo degli *attori* resta sicuramente una delle scelte più idonee, anche se i soggetti *naif* non sono affatto da disdegnare.

8. CONCLUSIONI

Nel presente lavoro sono state motivate e descritte le scelte metodologiche e le procedure di raccolta di un corpus mistilingue di parlato emotivo per l'italiano, il francese, l'inglese e il tedesco.

Sebbene i risultati riportati in questa sede si riferiscano nello specifico alla validazione del corpus raccolto, è stato comunque possibile accertare e verificare una serie di pre-supposti già confermati in precedenti ricerche, come ad esempio l'abilità da parte di giudici ascoltatori a riconoscere emozioni vocali presentate nella propria lingua madre e la maggiore riconoscibilità a livello vocale di emozioni come rabbia e tristezza (Scherer, Johnstone & Klasmeyer, 2003).

A parità di condizioni, l'utilizzo di *attori* ha rivelato una maggiore resa in termini di materiale utile, nonostante non siano da disdegnare i soggetti *naif*.

Rispetto al protocollo di raccolta delle emozioni vocali adottato in questa sede, va invece evidenziato come la produzione di frasi sulla base di scenari idonei (modalità di raccolta C nel presente lavoro) resti, al momento, una delle vie più percorribili nel caso di ricerche di tipo sperimentale (ancor più nel caso di ricerche di tipo cross-linguistico).

Mettendo a confronto i risultati della validazione effettuata, si rileva come la tipologia di test adottato, rispettivamente di identificazione delle emozioni nel T1 e di rappresentatività delle stesse nel caso di T2, possa diversamente discriminare tra produzioni utili o meno per via della diversa difficoltà cognitiva del *task*. Sebbene i due test risultino sufficientemente correlati, il diverso grado di difficoltà insito è senza dubbio da tenere in considerazione. Dal tipo di test e sulla base delle soglie di significatività individuate nel presente lavoro, dipende la caratterizzazione dei dati finali, come evidenziato, tra l'altro, nel grafico riportato in Fig. 9.

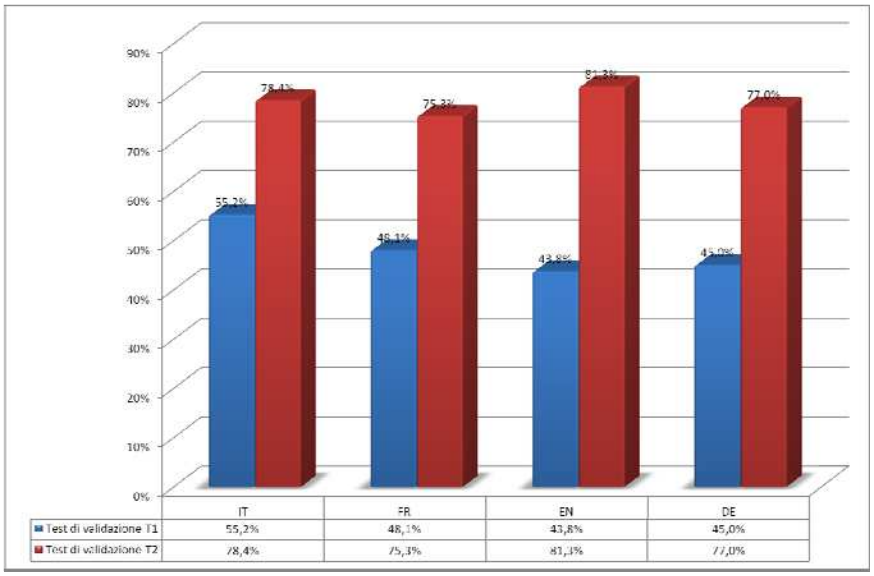


Figura 9: Confronto tra i due test di validazione (T1 e T2) con riferimento al materiale ritenuto utile per ciascuna lingua

In termini di capacità di discriminazione appare quindi evidente (vedi Figura 9) come sia sicuramente da prediligere (a scapito della quantità) un test di identificazione di emozioni che consenta di escludere tutte quelle produzioni che potrebbero creare confusione negli ascoltatori o nelle analisi acustiche che si intendono effettuare.

Tra gli obiettivi futuri, infine, a completamento dello studio qui intrapreso, sono attualmente in corso due ricerche parallele e complementari:

- un test percettivo di tipo cross-linguistico in cui vengono presentate a soggetti italiani produzioni emotive nelle quattro lingue qui raccolte;
- un'analisi acustica delle produzioni somministrate nel predetto test percettivo volta alla caratterizzazione acustica delle emozioni nelle quattro lingue.

RINGRAZIAMENTI

Un ringraziamento va a tutti coloro che si sono prestati nella raccolta e nella validazione delle registrazioni. Un ringraziamento va anche a Edwige Costanzo, Michael Cronin e Gudrun Wiesel per il loro aiuto nella traduzione degli scenari utilizzati per la raccolta del corpus.

9. BIBLIOGRAFIA

Abelin, Å. & Allwood, J. (2000), Cross linguistic interpretation of emotional prosody, in *SpeechEmotion 2000*, 110-113.

Abelin, Å. & Allwood, J. (2002), Cross linguistic interpretation of emotional prosody, *Papers in theoretical linguistics*, Gothenburg, 1-18.

Ambrus, D.C. (2000), Collecting and recording of an emotional speech database, *Technical Report*, Faculty of Electrical Engineering, Inst. of Electronics, Univ. of Maribor.

Anolli, L. & Ciceri, R. (1992), *La voce delle emozioni. Verso una semiosi della comunicazione vocale non-verbale delle emozioni*, Milano: Angeli.

Anolli, L., Wang, L., Mantovani, F. & De Toni, A. (2008a), La voce delle emozioni in giovani adulti cinesi e italiani, in *Comunicazione parlata e manifestazione delle emozioni,* Atti del 1° Convegno del Gruppo di Studio della Comunicazione Parlata, Padova, 29 novembre–1 dicembre 2004, (E. Magno Caldognetto, F. Cavicchio e P. Cosi, editors), Napoli: Liguori Editore, 2-44.

Anolli, L., Wang, L., Mantovani, F. & De Toni, A. (2008b), The Voice of Emotion in Chinese and Italian Young Adults, in *Journal of Cross-Cultural Psychology*, 39, 565-598.

Averill, J.R. (2004), Everyday Emotions: Let Me Count the Ways, *Social Science Information*, 43, 571-80.

Ax, A.F. (1953), The physiological differentiation between fear and anger in humans, *Psychosomatic Medicine*, 15, 433-442.

Baños, R., Liaño, V., Botella, C., Alcañiz, M., Guerrero, B. & Rey, B. (2006), Changing induced moods via virtual reality, in *Persuasive Tecnology for Human Well-Being: Setting the scene*, 3962, 7-15.

Barbaranelli, C. & D'Olimpo, F. (2007), *Analisi dei dati con SPSS. Vol. I: Le analisi di base*, Milano: LED.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russel, M. & Wong, M. (2004), 'You stupid tin box' - children interacting with the AIBO robot: a cross-linguistic emotional speech corpus, in *Proceedings of the 4th International Conference of Language Resources and Evaluation*, Lisbon, Portugal, May 26-28, 2009, 171-174.

Boersma, P. & Weenink, D. (2009), *Praat: doing phonetics by computer* [Computer program], retrieved from <http://www.praat.org/>.

Braun, A. & Oba, R. (2007), Speaking Tempo in Emotional Speech – a Cross-Cultural Study Using Dubbed Speech, in *Proceedings of the International workshop on Para-linguistic Speech – between models and data, ParaLing'07,* 3 August 2007, Saarbrücken, Germany, <http://www2.dfki.de/paraling07/papers/16.pdf>.

Breitenstein, C., Van Lancker, D. & Daum, I. (2001), The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample, *Cognition & Emotion*, 15(1), 57-79.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F. & Weiss, B. (2005), A database of German emotional speech, in *INTERSPEECH 2005*, 1517-1520.

Campbell, N. (2000), Databases of emotional speech, in *SpeechEmotion 2000*, 34-38.

Chung, S. J. (1999), Vocal expression and perception of emotion in Korean, in *Proceedings of the 14th International Conference of Phonetic Sciences*, San Francisco, USA, August 1-8, 1999, 969-972.

Chung, S. J. (2000), *L'expression et la perception de l'émotion extraite de la parole spontanée: évidences du coréen et de l'anglais,* Unpublished doctoral dissertation, Université de la Sorbonne Nouvelle, Paris III, France, <http://www.geocities.com/soojinchung/Finalthesis.pdf>.

Clark, D.M. (1983), On the induction of depressed mood in the laboratory: Evaluation and comparison of the Velten and musical procedures, *Advanced Behavior Research and Therapy*, 5, 27-49.

Douglas-Cowie, E., Campbell, N., Cowie, R. & Roach, P. (2003), Emotional speech: Towards a new generation of databases, *Speech Communication*, 40, 33-60.

Dromey, C., Silveira, J. & Sandor, P. (2005), Recognition of affective prosody by speakers of English as a first or foreign language, *Speech Communication*, 47, 351-359.

Eich, E., Ng, J. T. W., Macaulay, D., Percy, A. D. & Grebneva, I. (2007), Combining music with thought to change mood, in *The handbook of emotion elicitation and assessment,* (J. A. Coan, J. J. B. Allen, editors), London: Oxford University Press, 124-136.

Ekman, P. (1992), An argument for basic emotions, *Cognition and Emotion*, 6, 169-200.

Enos, F. & Hirschberg, J. (2006), A Framework for Eliciting Emotional Speech: Capitalizing on the Actor's Process, *LREC 2006 Workshop on Corpora for Research on Emotion and Affect*, Genova, Italy, May 23, 6-10.

Fairbanks, G. & Hoaglin, L.W. (1941), An experimental study of the durational characteristics of the voice during the expression of emotion, *Speech Monograph*, 8, 85-91.

Fairbanks, G. & Pronovost, W. (1939), An experimental study of the pitch characteristics of the voice during the expression of emotion, *Speech Monograph*, 6, 87-104.

Gerrards-Hesse, A., Spies, K. & Hesse, E.W. (1994), Experimental inductions of emotional states and their effectiveness: A review, *British Journal of Psychology*, 85, 55-78.

Gonzalez, G.M. (1999), Bilingual computer-assisted psychological assessment: an innovative approach for screening depression in Chicanos/Latinos, *Technical Report*, 39, Univ. Michigan.

Härtel, C.E.J. & Härtel, G. F. (2005), Cross-cultural differences in emotions: the why and how, *Social Science Information*, 44, 683-693.

Helfrich, H., Standke, R. & Scherer, K.R. (1984), Vocal indicators of psychoactive drug effects, *Speech communication*, 3, 245-252.

Iadarola, I. (2009), EMOVO, database di parlato emotivo per l'italiano, in *La Fonetica Sperimentale. Metodo e Applicazioni*, Atti del 4° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Arcavacata di Rende (CS), 3-5 dicembre 2007 (L. Romito, V. Galatà, e R. Lio, editors), Torriana: EDK Editore SRL, 293-323.

J.A. Coan & J.J. B. Allen (2007), *The handbook of emotion elicitation and assessment,* London: Oxford University Press.

Johnstone, T. & Scherer, K.R. (2000), Vocal Communication in Emotion, in *Handbook of Emotion* (M. Lewis & J. Haviland, editors), New York: Guilford, 220-235.

Johnstone, T., van Reekum, C.M., Hird, K., Kirsner, K. & Scherer, K.R. (2005), Affective speech elicited with a computer game, *Emotion*, 5, 513-518.

Kenealy, P. (1986), The Velten Mood Induction Procedure: A Methodological Review, *Motivation and Emotion*, 10(4), 315-335.

Kori, S. & Magno Caldognetto, E. (2003), La caratterizzazione fonetica delle emozioni: primi dati da uno studio cross-linguistico italiano-giapponese, in *Voce-Canto-Parlato. Studi in onore di Franco Ferrero* (P. Cosi, E. Magno Caldognetto & A. Zamboni, editors), Padova: Unipress,187-200.

Laukka, P. (2004), Vocal expression of emotion: discrete-emotions and dimensional accounts, *Ph.D thesis*, Uppsala University.

Magno Caldognetto E. & Kori S. (1983), Intercultural judgment of emotions expressed through voice, in *Quaderni del Centro di Studio per le Ricerche di Fonetica*, 2, 339-363.

Magno Caldognetto, E. (2002), I correlati fonetici delle emozioni, in *Passioni, emozioni, affetti* (C. Bazzanella & P. Kobau, editors), Milano: McGraw-Hill, 197-213.

Magno Caldognetto, E., Cavicchio, F., Cosi, P., Drioli, C. & Tisato, G. (2005), Parametri per lo studio delle modificazioni articolatorie del parlato emotivo, in *Misura dei parametri. Aspetti tecnologici ed implicazioni nei modelli linguistici*, Atti del 1° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Padova, 2-4 dicembre 2004 (P. Cosi, editor), Brescia: EDK Editore, 441-470.

Niedenthal, P.M., Krauth-Gruber, S. & Ric, F. (2006), *The Psychology of Emotion: Interpersonal Experiential, and Cognitive Approaches,* Principles of Social Psychology series, New York: Psychology Press.

Pavlenko, A. (2005), *Emotions and Multilingualism*, New York: Cambridge University Press.

Pell, M.D., Monetta, L., Paulmann, S. & Kotz, S.A. (2009), Recognizing emotions in a foreign language, *Journal of Nonverbal Behavior*, 33, 107-120.

Piot, O. (1999), Experimental study of the expression of emotions and attitudes in four languages, in *Proceedings of the 14th International Conference of Phonetic Sciences*, San Francisco, USA, August 1-8, 1999, 369–370.

Poggi, I. & Magno Caldognetto, E. (2004), Il parlato emotivo. Aspetti cognitivi, linguistici e fonetici, in *Atti del Convegno Italiano parlato*, Napoli, 14-15 febbraio 2003, (F. Albano Leoni, F. Cutugno, M. Pettorino, R. Savy, editors), Napoli: D'Auria Editore, CD-Rom.

Rottenberg, J., Ray, R. D. & Gross, J.J. (2007), Emotion elicitation using films, in *The handbook of emotion elicitation and assessment*, (J. A. Coan, J. J. B. Allen, editors), London: Oxford University Press, 9-28.

Sawamura, K., Dang, J., Akagi, M., Erickson, D. *et al.* (2007), Common factors in emotion perception among different cultures, in *Proceedings of the 16th International Conference of Phonetic Sciences*, Saarbrücken, Germany, August 6-10, 2007, 2113-2116.

Scherer, K.R. (1988), *Facets of emotion: Recent research*, Hillsdale, NJ: Erlbaum.

Scherer, K.R., Banse, R. & Wallbott, H.G. (2001), Emotion Inferences from Vocal Expression Correlate across Languages and Cultures, *Journal of Cross-Cultural Psychology*, 32, 76-92.

Scherer, K.R., Grandjean, D., Johnstone, L.T. & G. Klasmeyer, T. B. (2002), Acoustic correlates of task load and stress, in *Proceedings of the International Conference on Spoken Language Processing 2002*, Denver, Colorado, USA, September 16-20, 2017-2020.

Scherer, K.R., Johnstone, T. & Klasmeyer, G. (2003), Vocal expression of emotion, in *Handbook of the Affective Sciences* (D.J. Davidson, H. Goldsmith, H.R. Scherer, editors), New York/Oxford: Oxford University Press, 433-456.

Scherer, K.R., Banse, R., Wallbott, H. G. & Goldbeck, T. (1991), Vocal cues in emotion encoding and decoding, *Motivation and Emotion*, 15, 123-148.

Schröder, M. (2003), Experimental study of affect bursts, *Speech Communication*, 40, 99-116.

Schröder, M. (2004), Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis, *PhD thesis, PHONUS 7, Research Report of the Institute of Phonetics,* Saarland University.

Seibert, P. S. & Ellis, H. C. (1991), A convenient self-referencing mood induction procedure, *Bulletin of the Psychonomic Society*, 29, 121-124.

Shochi, T., Aubergé, V. & Rilliard, A. (2007), Cross-Listening of Japanese, English and French social affect: about universals, false friends and unknown attitudes, in *Proceedings of the 16th International Conference of Phonetic Sciences*, Saarbrücken, Germany, August 6-10, 2007, 2097-2100.

Sutherland, G., Newman, B. & Rachman, S. (1982), Experimental investigations of the relations between mood and intensive unwanted cognition, *British Journal of Medical Psychology*, 55, 127-138.

Thompson, W.F. & Balkwill, L.L. (2006), Decoding speech prosody in five languages, *Semiotica*, 158, 407-424.

Tickle, A. (1999), Cross-language vocalisation of emotion: methodological issues, in *Proceedings of the 14th International Conference of Phonetic Sciences*, San Francisco, USA, August 1-8, 1999, 305-308.

Tickle, A. (2000), English and Japanese speakers' emotion vocalization and recognition: A comparison highlighting vowel quality, *Speech and Emotion*, 104-109.

Velten, E. (1968), A laboratory task for induction of mood states, *Behaviour Research and Therapy*, 6, 473-482.

Ververidis, D. & Kotropulos, C. (2006), Emotional speech recognition: Resources, features, and methods, *Speech Communication,* 48, 1162-1181.

Wallbott, H. G. & Scherer, K. R. (1986), Cues and channels in emotion recognition, in: *Journal of personality and social psychology*, 51, 690-699.

Westermann, R., Spies, K., Stahl, G. & Hesse, F. W. (1996), Relative effectiveness and validity of mood induction procedures: a meta-analysis, *European Journal of Social Psychology*, 26, 557-580.

Williams, C.E. & Stevens, K.N. (1972), Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, 52, 1238-1250.

STABILITÀ DEI PARAMETRI NELLO *SPEAKER RECOGNITION*. LA VARIABILITÀ INTRA E INTER PARLATORE: F0, DURATA E *ARTICULATION RATE*

Luciano Romito [a], Rosita Lio [a], Pier Franceso Perri [b], Sabrina Giordano [b]
[a] Laboratorio di Fonetica, [b] Dipartimento di Economia e Statistica
Università della Calabria
luciano.romito@unical.it, lio.rosita@libero.it,
pierfrancesco.perri@unical.it, sabrina.giordano@unical.it

1. SOMMARIO

La tendenza della ricerca attuale in ambito di *Speaker Recognition* (SR) è volta a individuare informazioni quanto più oggettive possibili presenti nella voce umana analizzando la produzione di un parlatore senza occuparsi della sfera semantica, della produzione linguistica o della struttura sintattica e morfologica. In aggiunta i metodi noti come semiautomatici e parametrici si occupano di dati *considerati* statici. Tale scelta in primo luogo è giustificata dalla relativa facilità della misura e dal trattamento di un ristretto numero di parametri (cfr. Barlow & Wagner, 1998) e in secondo luogo perché la misura di dati statici è la naturale evoluzione di una tradizionale analisi linguistica (cfr. McDougall, 2006).

Sono i segmenti statici quelli utilizzati per lo studio delle lingue, si pensi agli inventari fonologici, alle aree di esistenza delle vocali costruite su porzioni stazionarie (*mid point* o *steady state*), alle rotazioni consonantiche o alle regole fonologiche. Tale analisi prende lo spunto dalla necessità di differenziare due lingue, due dialetti o una lingua da un dialetto. Così, grande spazio nelle riviste, occupano concetti quali isoglosse o isofone utilizzati per identificare confini ideali tra due lingue o tra due dialetti.

Quanto detto risulta funzionale per differenziare ma non per riconoscere, o addirittura identificare. Di fatto anche il concetto di isoglossa oggi viene sostituito dall'idea più 'analogica' di corridoio di transizione, una larga fascia dove coesistono variabili differenti che caratterizzano entrambe le lingue o i dialetti contigui.[1]

Un parlante nel produrre un messaggio o un atto comunicativo attraverso un meccanismo astratto (linguistico), organizza *target* e *goal* che, in seguito, verranno tradotti in azioni che si realizzeranno in un 'progetto fonetico'. Il meccanismo linguistico è essenzialmente l'insieme delle regole e della grammatica del parlante; è la lingua costituita dal lessico, dalla morfologia, dalle opposizioni fonologiche, dalla sintassi, ecc. Tale meccanismo è fortemente influenzato dall'età, dal sesso, dal controllo fonologico, da fattori sociali quali l'origine geografica, lo stato economico, il contesto, la scolarizzazione, ecc. Nolan, a tal proposito, nel 1997 (p.749) scrive: "In implementing the resources of their linguistic mechanism, speakers have to map them onto their individual anatomy. Whilst the requirements of communication may determine many of the details of speech articulation, we may hypothesize that there may be aspects of speech production where each individual

[1] In un paese della preSila catanzarese (Soveria Mannelli) coesistono due variabili per il passato e l'imperfetto: la variabile catanzarese [ˈji:vi] e quella cosentina [ˌsiɲuˈju:tu] 'sono andato'.

is free to find his or her own articulatory solution. The speaker's behavior here is not 'learned' as part of the shared knowledge of the linguistic community; rather it is acquired, probably by trial and error".

Due differenti parlanti possono eseguire progetti fonetici differenti per lo stesso scopo linguistico e le conseguenze acustiche di tali progetti possono aiutare molto nel differenziare, anche se, a nostro avviso le modifiche non riguarderanno la parte statica del segnale.

Lo scopo di questo progetto di ricerca, i cui primi risultati sono stati presentati ai Convegni AISV 2006 e 2007, è quello di studiare la variabilità interna di alcuni parametri acustici, di verificare la correttezza di un confronto o di una comparazione basata sul progetto fonetico e quindi su parametri dinamici e di comparare i risultati ottenuti con quelli basati su parametri considerati statici. Verrà analizzato soprattutto l'effetto prodotto da differenti canali di registrazione, da differenti stili di parlato e da differenti software di analisi. In questa ricerca l'attenzione non è focalizzata sul numero degli intervistati bensì sulla varietà dei canali di registrazione investigati e degli stili di parlato considerati.

2. PREMESSA

La voce è molto più di una semplice sequenza di suoni. Essa è intrinsecamente articolata e gran parte della sua complessità è legata ai rapporti tra le singole variabili che operano al suo interno come ad esempio il senso, il significato, le intenzioni, le emozioni, lo stato di salute, lo stato sociale, il livello di autostima, il livello di scolarizzazione, ecc. Tutto ciò, ovviamente, è molto importante dal punto di vista forense, almeno potenzialmente, visto che è difficile isolare acusticamente le variabili legate ad ogni singolo livello e considerato che tutte vengono veicolate in un unico canale: quello acustico. J. Laver a riguardo scriveva "The voice is the very emblem of the speaker, indelibly woven into the fabric of speech. In this sense, each of our utterances of spoken language carries not only its own message, but through accent, tone of voice and habitual voice quality it is at the same time an audible declaration of our membership of particular social and regional groups, of our individual physical and psychological identity, and of our momentary mood" (Laver, 1994: 2). Per effettuare una corretta misurazione quindi, è necessaria, una profonda conoscenza della correlazione esistente tra singola variabile ed effetto acustico prodotto. Solo grazie a questa competenza è possibile identificare i parametri da estrapolare, interpretare i segnali sonori e decidere se i campioni di voce sono comparabili.

La tendenza della ricerca italiana e del mondo occidentale in genere, in ambito di SR, è volta ad individuare informazioni quanto più possibili oggettive presenti nella voce umana. L'attenzione è quindi concentrata su tutte quelle informazioni acustiche presenti nella voce, trascurando, da una parte la sfera semantica, morfologica, fonologica e sintattica e dall'altra variabili stilistiche come quelle diafasiche, diastratiche o diatopiche presenti nel segnale sonoro.

Le indagini peritali fino ai primi anni '80 si basavano essenzialmente su aspetti strutturali (paradigmatici) più che su caratteristiche acustiche. Tali indagini linguistiche, glottologiche o sociolinguistiche, che diedero origine anche ad un filone di studio e di ricerca noto come 'sociolinguistica giudiziaria' (v. Trumper, 1979), furono utilizzate per identificare una comunità linguistica più che l'identità di un singolo parlante. Venivano utilizzate soprattutto in fase di indagini preliminari al fine di ricercare la provenienza di voci anonime presenti in rivendicazioni a sfondo terroristico o in casi di sequestro di persona. Tutto ciò era possibile soprattutto perché il materiale sonoro da analizzare e da studiare era abbon-

dante (si veda cosa è successo dopo l'applicazione del 'blocco telefonico'),[2] al contrario di quanto accade oggi (cfr. Romito *et al.*, 1996a, e 1996b). Il tempo dedicato ad ogni singola indagine (consulenza) dagli esperti, era molto maggiore e spesso frutto di stretta collaborazione tra competenze diverse quali la musica, la linguistica, la fisica acustica e la statistica.

Oggi sarebbe impossibile, e forse anche inutile, effettuare una consulenza (o perizia) linguistica; i motivi sarebbero da ricercare nella competenza dialettologica dell'esperto, nei tempi molto lunghi richiesti per l'analisi che poco si conciliano con la velocità imposta dalle indagini ed infine nella richiesta di consulenze che riguardano più il singolo parlante che la comunità di appartenenza; inoltre con materiale sempre più scarso sia in quantità che in qualità, sarebbe come attribuire un breve articolo anonimo pubblicato su un giornale ad uno scrittore noto.[3] Questo tipo di inchieste non si basano su ricorrenze lessicali o su intercalari, né tanto meno su minuzie o articolazioni particolari ma sull'individuazione di regole fonologiche, morfologiche e sintattiche presenti in una registrazione. Ad esempio, analizzando (sintatticamente) una registrazione anonima di un parlante meridionale calabrese potremmo concentrare la nostra attenzione sulla posizione che assume il pronome possessivo in alcuni particolari contesti come i nomi parentali. Potremmo constatare che i dialetti della costa calabrese Tirrenica (per esempio: Palmi, Delianuova) antepongono il pronome possessivo (*me patre*, *me frate*, *me soru* "mio padre, mio fratello, mia sorella") mentre al contrario i dialetti della costa calabrese Ionica (per esempio: Catanzaro lido, Soverato, Siderno, Locri) post pongono il pronome con diversi esiti di suffissazione (*patrimma*, *fratimma*, *soremma* o *sorma* "mio padre, mio fratello, mia sorella"). Ipotizzando di aver identificato nella registrazione anonima una provenienza del parlante dalla costa ionica calabrese, una seconda analisi (fonologica-fonetica) sempre sulla stessa registrazione potrebbe riguardare l'esito della doppia –LL- latina nei dialetti in questione. I dialetti del catanzarese prevedono un esito occlusivo retroflesso sonoro [ɖɖ] quindi ILLUM > *iɖɖu* "lui"; la zona più a sud sempre sulla costa Jonica come Roccella Jonica ecc. prevede un esito approssimante palatale sonoro [j] *iju* "lui" mentre ancora più a Sud (Badolato)

[2] Possibilità di rintracciare la provenienza di una telefonata, il numero e l'apparecchio telefonico. Tale operazione è possibile solo se la conversazione è sufficientemente lunga. Da allora i malviventi iniziarono a ridurre drasticamente la durata delle telefonate.

[3] Si pensi ad esempio alla grande difficoltà nel'attribuire alcuni articoli di giornale a Antonio Gramsci (cfr. Basile & Lana, 2009). C'è da aggiungere però che nonostante quanto affermato in Italia vengono effettuate senza alcun fondamento scientifico, delle perizie linguistiche finalizzate al riconoscimento del parlatore basate su poche frasi e su analisi basate su idioletto. È inutile affermare che anche se in Italia non esiste nessun controllo nelle Aule di Tribunale sul fondamento scientifico di alcune consulenze questo non significa che le stesse abbiano qualche valore. Negli Stati Uniti fino al 1997, l'ammissibilità del giudizio era essenzialmente basata sugli standard di Frye o di McCormik (Frye, 1923), dopo tale data vengono riportati i criteri essenziali per l'accettabilità di un metodo in ambito forense. I criteri per la scientificità di un metodo sono: qualunque teoria o tecnica che vuole essere tale deve essere testata; una tecnica deve essere stata pubblicata o sottomessa ad un *peer review*; deve essere stato considerato il potenziale errore; deve essere dichiarato se esiste uno standard e se questo è sotto il controllo dell'operatore della tecnica; il livello di accettazione della tecnica all'interno della comunità scientifica (cfr. Rose, 2002: 121).

l'esito è monovibrante alveolare sonoro [r] *iru* "lui". Individuata la zona si potrebbe andare ancora più nello specifico concentrandosi su processi metafonetici o su strutture sintattiche come l'uso dell'infinito opposto al */ma, ca*, *u/* + verbo al presente [poˈter(r)ə cjoˈviri] (base latina) versus [poˈtera ma ˈcjova] (base greco-bizantina) "potrebbe piovere". L'incrocio e la coesistenza di una serie di variabili (correttamente identificate) conduce alla identificazione di una precisa comunità linguistica. Tanto precisa sarà l'identificazione della comunità linguistica quanto unica e particolare risulterà essere la variabile considerata come nel caso dell'esito di –LL- in [ʎ] presente solo in un piccolo paese della Calabria aspro montana o ancora il processo di trittongazione riscontrato solo in un quartiere della città di Reggio Calabria.[4]

Ovviamente è cosa molto differente identificare un singolo parlatore attraverso la sola analisi linguistica.

3. EXCURSUS SUI METODI DI SR

I metodi di SR utilizzati oggi in ambito forense, vengono suddivisi in automatici, semiautomatici e soggettivi. In questa tassonomia, l'attenzione è rivolta all'intervento dell'operatore sull'analisi e sulla estrapolazione dei parametri utili alla comparazione. In questa sede si è scelto, invece, di basare la categorizzazione sui parametri utilizzati definiti 'statici', 'dinamici' e 'dinamico-selettivi'; un tentativo di avvicinare i metodi automatici con il controllo dell'operatore.

3.1 Metodi che utilizzano dati statici

In questa sezione possono essere sicuramente annoverati tutti i metodi semiautomatici e manuali definiti parametrici ed alcuni metodi automatici come quello basato sulla funzione dissipativa (*functional dissipation*).[5]

Per la maggior parte dei metodi semiautomatici e parametrici oggi utilizzati in ambito forense, i dati definiti statici sono identificati nelle porzioni stazionarie delle vocali (in genere quelle toniche). La scelta è giustificata dalla *relativa* facilità della misura e del trattamento di un ristretto numero di parametri (cfr Barlow & Wagner, 1998). Inoltre la misura delle parti stazionarie delle vocali toniche (dati statici) è la naturale evoluzione di una tradizionale analisi linguistica/dialettologica (cfr. McDougall, 2006), basti pensare al concetto già presentato di isoglosse e isofone o alle mappe tematiche basate su singole variabili o su inventari fonologici, e sistemi vocalici di derivazione latina (v., ad esempio, Tagliavini, 1982). Altro motivo invece riguarda la correlazione, anche questa di tradizione

[4] Si confrontino gli studi dialettologici a partire dagli anni '80.

[5] Il metodo attraverso la funzione dissipativa è ancora sperimentale e ad oggi non è mai stato utilizzato in ambito forense; cfr. Napoletani *et al.* (in stampa) per l'ambito forense e Napoletani et al. (2002) per l'ambito medico: "Functional dissipation is based on signal transforms, but uses the transforms recursively to uncover new features. We generate a variety of masking functions and 'extract' features with several generalized matching pursuit iterations. In each iteration the recursive process modifies several coefficients on the transformed signal with the largest absolute values according to the specific masking function; in this way the greedy pursuit is turned into a slow, controlled, dissipation of the structure of the signal that for some masking functions, enhances separation among classes".

linguistico-fonetica (cfr. Fant, 1960), tra l'impostazione articolatoria e il relativo effetto acustico (nel nostro caso il parametro da estrapolare). Così, ad esempio, il valore acustico della prima formante vocalica corrisponderà (con una relazione inversamente proporzionale) all'altezza della lingua lungo un asse basso-alto all'interno dell'apparato boccale, ecc. Nel nostro esperimento le misure vengono effettuate solo sulla porzione stazionarie delle vocali caratterizzate da accento frasale e meglio rispondenti al concetto di target articolatorio.

3.2 Metodi che utilizzano dati dinamici

I metodi che utilizzano esclusivamente dati (acustici) dinamici sono quelli automatici o semiautomatici.[6] Questi, molto diversi tra loro, si basano essenzialmente su spettri a lungo termine (LTS) o su coefficienti Melcepstrali e al momento non vengono utilizzati in ambito forense.[7] Esiste però la possibilità di associare a metodi che utilizzando dati statici un dato dinamico come la stima della velocità di eloquio o meglio di articolazione (*Articulation Rate*; cfr. Künzel, 1997; Zavattaro, 2005).

3.3 Altri metodi

Un discorso differente deve essere effettuato per i metodi soggettivi come i *voiceprints* (o confronto dei sonogrammi) e i metodi percettivi uditivi. Vengono entrambi utilizzati in ambito forense in Italia, nonostante la comunità scientifica internazionale ne sconsigli l'uso (soprattutto per il confronto dei sonogrammi) visto l'alta probabilità di errore. Per quanto riguarda il confronto dei sonogrammi in ambito forense (v. Tosi, 1979), la comunità scientifica si è più volte pronunciata sulla sua inaffidabilità (cfr. Gruber & Poza, 1995: 54-71). Tale metodo si basa essenzialmente su due protocolli: il primo protocollo è stato sviluppato da VIAAS (*Voice Identification and Acoustic Analysis SubCommittee*, della *International Association for Identification*) ed è stato pubblicato negli atti dell'associazione VCS 1991;[8] il secondo protocollo schematizzato dell'FBI è stato pubblicato in Koenig (1986: 2089-90).[9] I protocolli sono molto simili, entrambi sono soggettivi e basati

[6] Per un giudice, in maniera naturale e istintiva, è più facile comprendere il metodo uditivo o quello parametrico. Molto difficile è, invece, la comprensione del metodo automatico se non si ha una profonda conoscenza sia del metodo che dello speech processing.

[7] Un tentativo di correlare l'impostazione articolatoria e il coefficiente cepstrale è da imputare al lavoro di Clermont & Itahashi (1999), secondo cui la qualità vocalica (quindi i valori formantici) sono in stretta correlazione con la variazione del II e del III coefficiente cepstrale.

[8] VCS (1991:373-9): "Ideally, the exemplar should be spoken [by the suspect] in a manner that replicates the unknown talker, to include speech rate, accent, (whether real or feigned), hoarseness, or any abnormal vocal effect. In general, the suspect is instructed to talk at his or her natural speaking rate: if this is markedly different from the unknown sample, efforts should be made through recitation to appropriately adjust the speech rate of the exemplar. Spoken accents or dialects, both real and feigned should be emulated by the known speaker. If any other unique aural or spectrally displayable speech characteristics are present in the questioned voice, then attempts should be made to include them in the exemplars".

[9] "AFTI: Visual comparison of spectrograms involves, in general, the examination of spectrograph features of like sounds as portrayed in spectrograms in terms of time, frequency and amplitude… Aural cues… include resonance quality, pitch, temporal factors,

sull'esperienza dell'esperto. Le critiche mosse a tale metodo riguardano l'identificazione degli elementi minimi utilizzati per la comparazione (Hollien, 1990: 215), l'impossibilità di presentare le evidenze dell'esaminatore o le caratteristiche numerabili e, infine, l'utilizzo di parametri qualitativi (Aitken, 1995: 14-15). Al momento il metodo sembra essere più intuitivo che analitico.[10]

Il metodo Percettivo-Uditivo sfrutta la capacità del singolo individuo a riconoscere la similitudine o la differenza tra due voci. Alcuni tra i metodi utilizzati sono; il *Panel Approach* (comparazione di coppie di frasi anche di diversa durata e tipo; le risposte sono in percentuale e si basano su caratteristiche stilistiche, linguistiche e acustiche); il *Direct Processing* (un ascoltatore esperto ascolta un intero brano e ne identifica la voce) e *l'Aural-Perceptual Approach* o *Aural-Spectographic Method*[11] (che prevede una combinazione del Metodo Percettivo-Uditivo e del confronto dei *Voiceprints* o sonogrammi; cfr. Hollien, 1990: 215; McDermott *et al.*, 1996).[12] Il metodo è quello di più facile comprensione per un giudice e una Corte.

4. PARAMETRI STATICI E PARAMETRI DINAMICI

Un parlante nel produrre un atto linguistico mette in campo tutta una serie di aggiustamenti e di processi coarticolatori sia segmentali che sovrasegmentali. Ogni produzione è influenzata oltre che da semplici processi fonologici e fonetici anche da variabili diafasiche, diastatiche, diatopiche, come già detto. Ad esempio, il tempo e la velocità di eloquio sono variabili che influenzano ed incidono molto sulla produzione del parlato spontaneo. Senza entrare molto nello specifico le lingue del mondo vengono classificate e divise proprio in base alla gestione del tempo. In tutte le lingue aumentare la velocità vuol dire ridurre la precisione nell'articolazione di alcuni segmenti o sillabe o addirittura cancellarne alcune ritenute ridondanti per la comprensione del messaggio linguistico. Quindi, mentre da una parte in molti dialetti meridionali si registra la sola

---

inflection, dialect, articulation, syllable grouping, breath pattern disguise, pathologies and other peculiar speech characteristics."

[10] V. Kersta (1962: 1253): "Voiceprint identification is a method by which people can be identified from a spectographic examination of their voice. Closely analogous to fingerprint identification, which uses the unique features found in people's fingerprints, voiceprint identification uses the unique features found in their utterances". V. anche Nash, citato in Hollien (1990: 224): "As each one of the ridges of your fingers or on the palm of your hand differ from each other, so do all of the other parts of your body. They are unique to you … including your voice mechanism". Tale metodo è stato sviluppato e commercializzato da Kersta (1962). Infine, cfr. Tosi (1979): "... the legal application of speaker identification, which at present still consists mainly in the practice of visual examination of spectrograms...".

[11] Questo metodo è ancora usato almeno fino al 2001 dall'FBI (cfr. Nakasone & Beck, 2001), dalla Polizia Giapponese (cfr. Osanai, 1995), in Israele, Italia, Spagna, Columbia (cfr. Rose, 2002), non viene più usato in Olanda e Germania (cfr. Künzel, 1994: 138).

[12] Sentenza dello Stato della California: "That the aural spectrographic analysis of the human voice for the purposes of forensic identification has failed to find acceptability and reliability in the relevant scientific community, and that therefore, there exists no foundation for its admissibility into evidence in this hearing pursuant to the law of California".

centralizzazione delle vocali finali (ipoarticolazione) e quindi parole diverse come *i chiodi*, *il chiodo* e *piove* possono essere prodotte indistintamente [‘cjovə] perché comunque ci penserà il contesto ad esplicitare e differenziare le produzioni, in altre lingue la velocità si ottiene sia centralizzando che cancellando sillabe ritenute ridondanti. Così una frase francese come *je ne sais pas* può, se prodotta velocemente diventare [jənesɛ’pa] > [jə̃sɛ’pa] > [ʃɛ’pa] fino a raggiungere la sua massima influenza e minima produzione in [‘ʃpa]. Le produzioni oscillano quindi da una iperarticolazione dove tutta l’informazione è veicolata dal segnale, ad una ipoarticolazione dove invece l’informazione è data dal contesto o dalle conoscenze pregresse. Come si può notare, gli elementi che non vengono mai intaccati o indeboliti sono i segmenti Tonici (sia vocali che sillabe), quelli definiti Target, Goal e quelli all’interno dei quali troveremo la parte stazionaria da misurare.

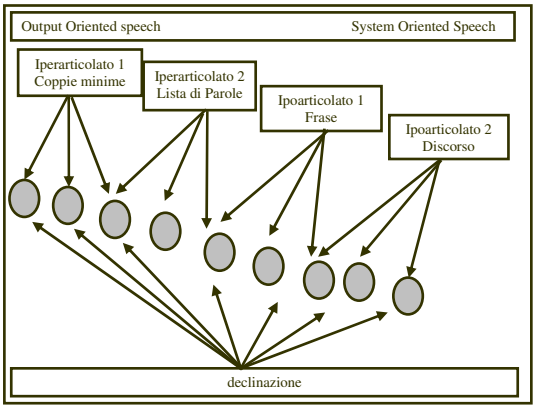


Figura 1: Grafico adattato da Romito *et al.*, 1997

La vocale (o sillaba) tonica è, quindi, (almeno teoricamente) l’unico elemento che raggiunge il target articolatorio, meno intaccato e influenzato dalla coarticolazione, sempre presente e mai cancellata e più vicino a quella ‘idea’ acustica che abbiamo nella testa e nella *langue* degli strutturalisti. Durante la produzione di questi elementi, ritenuti massimi portatori di informazioni linguistiche, fisiologicamente assistiamo al ‘congelamento’ delle posizioni geometriche e dei volumi creati all’interno dell’apparato boccale. Tale porzione temporale viene definita *steady state*, *mid point* o semplicemente ‘parte stazionaria’. In queste porzioni le formanti hanno un andamento lineare stabile e soprattutto (quasi) parallelo. Questi dati acustici definiti ‘stabili’ (da noi definiti ‘statici’), vengono considerati, nei metodi di comparazione, rappresentativi di ogni singolo parlante. Tale concezione teorica non tiene conto delle aree di esistenza vocaliche, della qualità fonetica nonché della qualità personale introdotta da Ladefoged.[13] Le differenti vocali vengono identificate e riconosciute dall’ascoltatore perché i valori delle formanti ricadono nell’area di esistenza ‘propria’ di quella vocale. Tali valori, quindi, così come le impostazioni articolatorie correlate, sarebbero, sempre (teoricamente), rappresentative più della lingua parlata che del

[13] Qualità personale e qualità fonetica: tutte le vocali che ricadono nella stessa area di esistenza hanno uguale qualità fonetica e diversa qualità personale se prodotte da differenti persone.

singolo parlante (cfr. Romito *et al.*, 1996, 1997; ma v. soprattutto Lindbloom, 1990). La mutua comprensione tra appartenenti alla stessa comunità linguistica è possibile proprio grazie, alla condivisione delle stesse aree di esistenza e quindi degli stessi target articolatori.[14]

5. IL PROGETTO FONETICO

Un parlante nel produrre un messaggio o un atto comunicativo, attraverso un meccanismo astratto (sul piano linguistico), organizza, target e goal che in seguito verranno tradotti in azioni che si realizzeranno in un 'progetto fonetico'. Come già detto, il meccanismo linguistico è essenzialmente l'insieme delle regole e della grammatica del parlante; è la lingua – costituita da lessico, morfologia, opposizioni fonologiche, sintassi, ecc. – ed è fortemente influenzato dall'età, dal sesso, dal controllo fonologico, da fattori sociali, dalla provenienza geografica, dallo stato economico, dal contesto e dalla scolarizzazione.

Due differenti parlanti possono attuare progetti fonetici differenti per lo stesso scopo o *target* linguistico e le conseguenze acustiche di tale progetto possono aiutare molto nel differenziare i parlanti ma, fatto estremamente importante non interesseranno la parte statica del segnale (cioè la qualità linguistica).

Lo scopo o il *target* può essere comune e avere caratteristiche linguistiche comuni, al contrario il progetto e il percorso che conduce (o guida) allo scopo o al *target*, è individuale e personale. In questa sede il percorso viene chiamato 'progetto fonetico'. Ipotizziamo inoltre, che il progetto fonetico si differenzi maggiormente nei dati dinamici di quanto non faccia nei dati statici e che le transizioni siano più individuali e personali di quanto non lo siano le parti stazionarie (con valore linguistico). Questo lavoro vuole verificare se è più corretto effettuare un confronto o una comparazione (cfr. Rose, 2002; ma anche Ezzaidi, Rouat & O'Shaughnessy, 2001), utilizzando il 'progetto' e i dati 'dinamici' o il metodo parametrico e i dati statici. Lo scopo di questo lavoro, è quindi quello di comparare i dati statici con i dati dinamici, in seguito verranno anche effettuate sperimentazioni sul confronto dei dati acustici articolatori (formanti, F0 ecc.) con dati più prettamente acustici (MFCC, LTS, ecc.). Nel prossimo futuro pensiamo di comparare solo ed esclusivamente le transizioni.

6. MATERIALI E METODI

Lo scopo di questo progetto di ricerca è quello di studiare la stabilità di alcune variabili definite statiche, dinamiche e dinamico-selettive.

In questa sede, ci occuperemo della frequenza fondamentale (F0) e della dimensione temporale nel parlato attraverso lo studio di registrazioni avvenute attraverso differenti canali e differenti modalità di produzione. Gli esperimenti si basano (come anche quelli relativi a Romito *et al.* (2007, 2008) su una selezione del *corpus* di voci intercettate PRIMULA.[15]

---

[14] Rammentiamo che la qualità vocalica cioè la posizione sullo spettro delle diverse formanti viene definita qualità linguistica.

[15] Cfr. http://www.linguistica.unical.it/labfon/home_corpus_primula.html e il paragrafo successivo per una maggiore definizione.

Le ipotesi di partenza presuppongono (cfr. Romito & Lio, 2008) che una variabile sia portatrice di informazione in ambito di SR quando:

a) mostra una alta variabilità interparlatore e una bassa variabilità intraparlatore;
b) è resistente al camuffamento;
c) ha una alta frequenza di occorrenza;
d) è robusta durante la trasmissione;
e) è relativamente facile da identificare e misurare.

6.1 Scelta del Campione

PRIMULA è un corpus ristretto di voci calabresi ideato e creato presso il Laboratorio di Fonetica dell'Università della Calabria per la valutazione delle metodologie e dei sistemi di riconoscimento del parlatore con particolare attenzione all'ambito forense. Il corpus costituisce una base comune sulla quale misurare le tecniche e le metodologie utilizzate in ambito di SR. Ciò che caratterizza PRIMULA è la modalità con cui lo stesso è stato creato. Durante la fase di ideazione del *corpus* e successivamente durante la fase di acquisizione dello stesso si è ritenuto di dover simulare una situazione reale al fine di avere, a prodotto finito, situazioni simili o quantomeno assai vicine a quelle che si presentano di norma nella maggior parte dei casi forensi. Proprio in virtù di ciò una parte delle registrazioni è stata effettuata con attrezzature normalmente utilizzate per le intercettazioni (grazie all'ausilio di alcuni Commissariati di Polizia e ditte private normalmente utilizzate nelle fasi di intercettazione dagli organi inquirenti). È stato così possibile registrare contemporaneamente ed in parallelo lo stesso materiale prodotto attraverso una microspia installata su un'autovettura e attraverso un cellulare collegato con un telefono fisso. Tale materiale registrato costituisce un'intercettazione 'ambientale' (in automobile) e una registrazione telefonica (tra utenza cellulare e utenza di rete fissa). Il corpus contiene poi, una serie di registrazioni di 'controllo' in condizioni differenti.

Il *corpus* PRIMULA consta di oltre 900 files raggruppati in differenti tipologie di registrazione: camera silente (segnale di alta qualità, utilizzato come training test o come saggio fonico); telefonata in ambiente rumoroso, in strada e in auto; infine intercettazione ambientale in auto (in movimento, durante una sosta e fuori dall'auto). Il tutto per 4 parlatori di sesso maschile con tipologie di produzione differenti sia di parlato letto (singole frasi lette per dieci volte e lettura di singole frasi), parlato spontaneo (sia in dialetto calabrese che italiano regionale) ed inoltre con tre diverse tipologie di voce: voce alta,[16] voce normale e voce bassa.[17]

6.2 Scelta dei Parametri

6.2.1 Parametri acustici: frequenza fondamentale (F0)

I parametri scelti sono stati differenziati in 'statici', 'dinamici' e 'dinamico-selettivi'. La differenza risiede, non certo nella variabile, ma nella sua misurazione. I parametri statici comprendono le misurazioni medie di F0 effettuate nella porzione stazionaria della vocale tonica con accento primario con una finestra di analisi non inferiore a 20 ms. Ciò rende

[16] Per ottenere una voce alta in maniera naturale, il parlante ha letto le frasi oggetto del test con una cuffia che diffondeva musica con un preciso valore di dB.
[17] Per voce bassa si è preferito una voce mormorata.

relativamente semplice la misurazione ma per ottenere un numero sufficiente di dati estrapolati è necessario avere un segnale molto lungo.

Per i parametri dinamici invece è stato considerato l'andamento globale della F0 con valori misurati ogni 0,01 sec. (i valori vengono estrapolati automaticamente su tutto il segnale: vocali, approssimanti, glides e consonanti sonore); l'analisi è molto veloce e si riesce ad ottenere un gran numero di dati anche su un segnale molto breve.

Per i parametri dinamico-selettivi è stata misurata la F0 in maniera dinamica ma solo ed esclusivamente su porzioni vocaliche (siano esse toniche che atone). In questo ultimo caso le vocali sono state differenziate sotto il profilo sia percettivo che spettrale in vocali toniche (VT), divise in *bad* (BVT), cioè vocali toniche che hanno subito un processo di deaccentuazione o caratterizzate da accento frasale secondario e *good* (GVT), cioè vocali sicuramente toniche con accento primario, e vocali atone (VA), suddivise anch'esse in *bad* (BVA) cioè vocali atone finali di parola o di frase e *good* (GVA) cioè vocali che nonostante lo status fonologico di atone vengono percepite come vocali qualitativamente con buona dispersione sul quadrilatero vocalico e buona rappresentazione spettrale. Anche su un segnale relativamente breve è possibile rilevare un sufficiente numero di dati. In un secondo è possibile trovare da 5 a 7 elementi vocalici.

6.2.2 Parametri linguistici: la velocità di eloquio e l'Articulation Rate

Il valore dell'*Articulation Rate* (AR) è stato misurato, sempre sullo stesso materiale sonoro, secondo la seguente formula:[18]

$$AR = \frac{\text{numero di sillabe fonetiche}}{\text{durata della catena fonica}}$$

dove per 'catena fonica', gruppo di respiro (o *run*) si intende tutto ciò che è presente tra due pause. Inoltre, in accordo con Künzel (1997) e Zavattaro (2005: 30), ai fini dello SR non vengono utilizzate le catene foniche con un numero di sillabe inferiore a 6. È stato infatti dimostrato che in questi casi la variabilità interna aumenta enormemente (Zavattaro, 2005).

6.3 Scelta degli algoritmi e dei metodi

Per quanto riguarda gli algoritmi scelti (soprattutto per la misura dei dati statici) la nostra attenzione si è soffermata su alcuni software normalmente utilizzati in ambito forense per l'analisi del segnale e l'estrapolazione dei parametri formantici: IDEM[19] (soprattutto nella sua sezione chiamata *Ares*), *Praat*[20] e *Multi-Speech*[21].

[18] Per una discussione completa sull'*Articulation Rate*, nonché sulle diverse definizioni presenti in letteratura si veda Romito et al., (2006).

[19] IDEM è un software per l'analisi del segnale e per l'identificazione del singolo parlante. È stato sviluppato dalla Fondazione Ugo Bordoni in collaborazione con l'Arma dei Carabinieri. Il software è composto da tre differenti moduli: *Ares* per la misura e l'estrapolazione dei dati parametrici, *Edit* per le funzioni di editing del segnale e *Spread* per la comparazione statistica dei dati. Tale software utilizza contemporaneamente l'analisi LPC e l'analisi cepstrale rappresentati su una finestra di analisi FFT. Il numero dei coefficienti e la finestra di analisi è stata mantenuta fissa, la ricerca della porzione da analizzare è manuale.

[20] *Praat* – che in olandese significa "parola" o "parlare" – è un *software open source* per l'analisi del segnale. È stato sviluppato da Paul Boersma e David Weenink dell'Università

6.4 Le analisi statistiche

Ci preme innanzitutto sottolineare che le analisi statistiche effettuate in questo lavoro non mirano all'identificazione del parlatore in senso forense, ma esclusivamente alla comparazione di campioni di dati al fine di misurare la variabilità intra e inter parlatore.[22] Per l'analisi descrittiva dei dati ci avvarremo di misure di sintesi, dell'ausilio grafico e dell'indice di affidabilità α di Cronbach, mentre per il confronto tra medie si farà uso dell'ANOVA ad una e a due vie e dell'ANOVA multivariata o MANOVA.[23]

7. LEGENDA E TABELLE

Le tabelle, che verranno presentate, conterranno i risultati dei confronti effettuati tra le medie di F0 o di *Articulation Rate* al variare dei parlanti, dei canali degli stili e dei segmenti. Il livello di significatività è sempre fissato a 0,05.

La legenda seguente favorisce la lettura delle tabelle statistiche che verranno presentate nei §§ successivi e l'individuazione delle variabili di volta in volta utilizzate.

I parlanti maschili sono 4, tutti meridionali e provenienti da 4 province differenti. Saranno segnalati attraverso le seguenti etichette LR, SC, AM e VG.

I canali e gli stili utilizzati sono i seguenti:

Micro LS = Canale: *Ambientale in auto - Microspia*; Stile: *lettura*;

Micro PS = Canale: *Ambientale in auto - Microspia*; Stile: *Parlato Spontaneo*;

Micro PS out = Canale: *Ambientale fuori dall'auto - Microspia*; Stile: *Parlato spontaneo*;

Tel Auto LS= Canale: *in Auto- Telefonico*; Stile: *lettura*;

Tel Strada LS = Canale: *in Strada- Telefonico*; Stile: *lettura*;

Tel Aula LS = Canale: *in Aula- Telefonico*; Stile: *lettura*;

Lezione = Canale: *in Aula-microfono*; Stile: *spontaneo formale*.

---

di Amsterdam e viene presentato nell'intro del *software* come "a computer program with which you can analyze, synthesize and manipulate speech".

[21] *Multi-Speech, Model 3700*, è un *software* per Windows che permette di campionare, analizzare e ascoltare. È completo di simboli IPA, e permette di editare, estrarre il pitch ed effettuare analisi formantiche attraverso LPC o FFT. È stato prodotto e sviluppato dalla *Kay Elemetrics Corp*. e dalla *Speech Technology Research Ltd.*.

[22] In ambito di *speaker recognition* (anche se al momento in Italia non esiste un protocollo standard), le variabili utilizzate per l'identificazione del parlatore sono esclusivamente acustiche e prevedono la misura media della frequenza fondamentale (F0) e delle frequenze formantiche su segmenti vocalici. I risultati vengono forniti attraverso il *likelihood ratio*. In alcuni casi viene anche fornito la stima dell'errore di falsa identificazione e di mancato riconoscimento utilizzando una popolazione di riferimento.

[23] Per applicare la metodologia dell'ANOVA bisogna verificare che vengano soddisfatte le ipotesi su cui poggia: normalità (verificata con il test di Smirnov o Shapiro) e omogeneità delle varianze (attraverso il test di Lavene. In quest'ultimo caso sono stati scelti test robusti per varianze non omogenee come Welch e Brown – Forsythe e Tamhane nel caso in cui l'omogeneità non venga soddisfatta. Le tabelle, sia relative alla normalità che alla omogeneità della varianza, per motivi di spazio non verranno presentate ma verrà segnalata di volta in volta nella tabella il test post hoc utilizzato.

Altra variabile è la porzione di segnale utilizzata per la misurazione e la stima della frequenza fondamentale. Essa è infatti stata misurata sui seguenti segmenti fonici:

VTG = Vocali Toniche ritenute *Good* sia percettivamente che spettrograficamente;

VTB = Vocali Toniche ritenute *Bad* sia percettivamente che spettrograficamente;

VAG = Vocali Atone ritenute *Good* sia percettivamente che spettrograficamente;

VAB = Vocali Atone ritenute *Bad* sia percettivamente che spettrograficamente.

8. CONTROLLO DEI DATI

8.1 Variabile Metodo-Algoritmo

La prima comparazione effettuata riguarda i dati estrapolati sullo stesso materiale attraverso diversi *software* con diversi algoritmi (per le specifiche relativa alle misure con i singoli *software* e alla segmentazione dei segnali si veda Romito & Galatà, 2008).

Parlanti, Canali e Stili	IDEM	Multispeech	Praat
LR Micro Letto	157.78	156.36	158.80
LR Micro PS in	142.30	144.20	145.93
LR Micro PS out	188.42	189.16	190.14
LR Tel Auto Letto	158.17	159.69	164.85
LR Tel Strada Letto	132.44	133.72	134.25
VG Micro PS	162.22	163.87	174.11
VG Tel Auto Letto	144.30	143.94	150.64
AM Tel Auto Letto	152.9	150.07	152.30
AM Micro PS	151.11	151.30	146.16
SC Micro Letto	124.31	128.00	125.32
SC Tel Auto Letto	152.15	150.88	154.95

Tabella 1: Valori medi di F0 statico misurato solo sulle vocali toniche per parlanti (canale e stile) e *software*

L'analisi scelta è quella multivariata (MANOVA); la variabile utilizzata è la frequenza fondamentale, i parlanti sono LR, SC, AM e VG (divisi anche per canale e stile). Le misure sono state effettuate con i seguenti software: *Idem, Praat* e *Multi-Speech*.

Il risultato della comparazione tra le medie della frequenza fondamentale ottenute per i diversi metodi-algoritmi dimostra che non vi è alcuna differenza (Sig=0,7873).[24] Se ne deduce che il metodo-algoritmo non influenza la misura di F0; ciò rende possibile comparare campioni misurati con metodi-algoritmi differenti.[25]

Anche il risultato ottenuto con la stima dell'indice di affidabilità (α di Cronbach) conferma quanto precedentemente affermato; il risultato di 0,980 confrontato con la tabella di riferimento presente in George & Mallery (2003: 231) rileva che il giudizio di affidabilità può essere definito *Excellent*.

[24]La differenza risulta essere altamente significativa per un valore di *Sig* <0.01, moderatamente significativa per un valore di *Sig* compreso tra 0.05 e 0.01 e non significativa per valori di *Sig* > 0.05.

[25] Ricordiamo che i risultati riguardano esclusivamente la misura del valore della frequenza fondamentale. Molto differenti sono, invece, i risultati relativi alle frequenze formantiche che presenteremo in un prossimo lavoro.

Valori	Giudizi di affidabilità
0.9	Excellent
.> 0,8	Good
> 0,7	**Acceptable**
> 0,6	**Questionable**
> 0,5	Poor, and
< 0,5	Unacceptable

Tabella 2: Giudizi di affidabilità secondo George & Mallery (2003:231)

8.2 Variabile 'canale'

Abbiamo testato l'affidabilità della variabile canale (telefono, ambiente rumoroso, auto ecc.) attraverso la stima dell'α di Cronbach. L'ipotesi formulata è la seguente: quanto la differenza del canale utilizzato influenza la misura della frequenza fondamentale?

I risultati mostrati nella tabella 2 rivelano un giudizio *questionable* associato all'indice di affidabilità per i parlanti LR (α = 0.592) e SC (α = 0.576), mentre il livello diventa *acceptable* per il parlante VG (α = 0.722). Le misure effettuate sullo stesso materiale sonoro, con gli stessi algoritmi sullo stesso parlante, vengono parzialmente influenzate dal canale utilizzato per la registrazione. Quanto affermato è solo una controprova del fatto che i dati ottenuti da misurazioni di materiale sonoro registrato su canali differenti non devono essere comparati tra di loro vista la difficoltà nello stabilire se la eventuale differenza riscontrata sia da attribuire al parlante o al canale di registrazione.

8.3 Distribuzione dei dati

Un ultimo controllo riguarda gli stili di voce utilizzati. Bisogna premettere che la frequenza fondamentale è il correlato acustico della vibrazione delle corde vocali e che il mormorio, il bisbiglio, il sussurro, e in genere la voce prodotta con bassa intensità, prevedono un'assenza di tale vibrazione (o almeno non completa; in alcuni casi infatti, le corde non vibrano per tutta la loro lunghezza vista anche la presenza di un'apertura attraverso le cartilagini aritenoidee che rende difficile la realizzazione del processo di Bernoulli).

La distribuzione dei dati, presentata di seguito, è relativa alla stessa frase, letta dallo stesso parlante, misurata con lo stesso algoritmo, nello stesso ambiente (camera silente), ma con livelli di voce differente: alta, normale e bassa. Come si evince dai grafici di seguito presentati, lo stile di voce influenza molto le misure della frequenza fondamentale. Innanzitutto nel caso di voce bassa il numero delle occorrenze è ridotto (come era prevedibile), i dati non hanno una distribuzione normale ed inoltre il valore medio e la mediana sono molto più bassi rispetto a quelli relativi alla voce normale e della voce alta (che invece innalza i valori di F0).

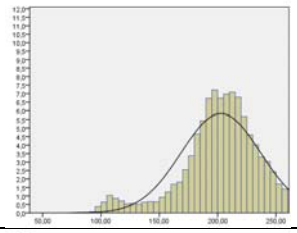
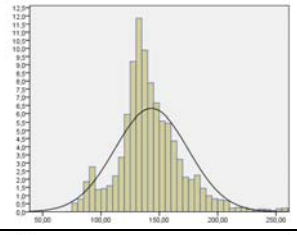
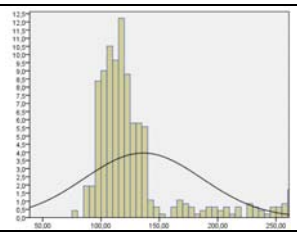
Modalità di Registrazione	Distribuzione dei dati
Voce Alta	
Voce Normale	
Voce Bassa	

Figura 2: Curve di normalità con identica scala delle misure in Hz della F0 relative allo stesso parlatore sulle stesse frasi in camera silente con Voce Alta, Normale e Bassa[26]

Nel caso della comparazione delle medie di F0 estrapolate dalle registrazioni effettuate su diversi canali e da diversi parlanti si è deciso di omettere lo stile di voce basso. Lo stesso stile verrà, invece, considerato nella comparazione della *Articulation Rate*, in quanto la mancanza di vibrazione delle corde vocali non influenza fisiologicamente la velocità di eloquio.

9. ANALISI DEI DATI STATICI

Effettuati i test di controllo sui dati e sui metodi di estrapolazione degli stessi, concentriamo la nostra attenzione sulla variabilità interna di ogni singolo parlante in funzione dei canali e degli stili di parlato.
Verranno confrontati i valori della frequenza fondamentale estrapolati da 4 parlanti (LR, SC, VG e AM) in funzione dei diversi canali (registrazione ambientale, in auto e fuori dall'auto, tramite telefono in strada e in auto). Contemporaneamente verranno considerate le registrazioni in funzione dello stile di parlato (lettura di frasi e parlato spontaneo). Il materiale in nostro possesso non è uniforme quindi anche i confronti effettuati all'interno di

[26] La scala è stata volutamente uniformata anche se ciò ha causato il taglio della coda destra nel caso di voce alta.

ogni singolo parlante saranno differenti. Nelle tabelle seguenti verranno riportate alcune misure statistiche calcolate sui *dati statici* di ogni parlante (le tipologie di registrazione non sono uguali per tutti i parlanti analizzati).

	N	Media	d std.	Min	Max
LR micro LS	59	157,92	29,223	103	216
LR Micro PS in	17	158,18	28,432	114	211
LR Micro PS in(2)	20	144,30	20,055	113	211
LR Tel Strada LS	20	152,90	36,336	103	222
Lr Micro PS out	17	151,12	31,774	99	205
LR Tel Auto LS	20	152,15	32,300	104	211
Totale	153	154,00	29,745	99	222

Tabella 3: Misure di F0 per LR

	N	Media	D. std.	Min	Max
SC Tel Auto LS	63	142,30	30,238	98	216
SC Micro LS	59	142,32	29,966	99	216
Totale	122	142,31	29,982	98	216

Tabella 4: Misure di F0 per SC

	N	Media	D. std.	Min	Max
VG Tel Auto LS	21	188,43	29,072	129	242
VG Micro PS	18	162,22	26,559	136	216
Totale	39	176,33	30,587	129	242

Tabella 5: Misure di F0 per VG

	N	Media	D. std.	Min	Max
AM Micro PS	25	132,44	19,744	90	174
AM Tel Auto LS	19	124,32	25,684	91	170
Totale	44	128,93	22,590	90	174

Tabella 6: Misure di F0 per AM

Al fine di verificare la presenza di un effetto ‘canale-stile’ sulla frequenza fondamentale si utilizza un’analisi ANOVA ad una via (F0 *by* ‘canale-stile’) per ciascun parlante.

L’analisi ANOVA fornisce per il parlante LR un valore della statistica F pari a 0.743 con Sig = 0.592. Questo indica che la variabilità interna al parlante non è molto elevata e

quindi diversi canali e stili non provocano differenze significative nei valori medi di F0. Analoghe conclusioni valgono per i parlanti SC (F = 0.00 e Sig = 0.997) e AM (F = 1.410 e Sig = 0.242).

Unico dato contrastante riguarda, invece, il parlante VG per il quale si è riscontrato un effetto rilevante del canale-stile sulla frequenza fondamentale (F = 8.524) evidenziando una significativa variabilità intraparlatore.

Il confronto in questo caso ha riguardato sia i canali che gli stili differenti. È stata comparata la media dei valori estrapolati da una registrazione avvenuta in auto per mezzo telefono con uno stile 'lettura frase' con una registrazione avvenuta in auto attraverso microspia con uno stile 'parlato spontaneo'.

Effettuata questa prima comparazione e questo controllo sulla variabilità interna ad ogni singolo parlatore in funzione del canale e dello stile, abbiamo successivamente comparato i valori medi delle F0, definite statiche ed estrapolate attraverso i diversi canali e i differenti stili (per i dati descrittivi si faccia riferimento alla tabella 1), per i diversi parlatori. A tal fine l'analisi ANOVA fra i gruppi (LR, SC, VG e AM) evidenzia una differenza significativa tra i valori medi di F0 registrati per i diversi parlanti come mostra la tabella seguente:

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	66777,298	11	6070,663	7,241	**,000**
Within Groups	290066,032	346	838,341		
Total	356843,330	357			

Tabella 7: ANOVA per F0 *by* parlante

Il passo successivo ha riguardato l'analisi puntuale di ogni singola comparazione, incrociando gli stili dei parlatori come ad esempio **LR** Micro LS vs **SC** *Tel Auto LS*, o **LR** *Micro LS* vs **SC** *Micro LS.*[27]

Dalle analisi effettuate risulta che in molti casi la differenza non è significativa. Questo implica che il valore di F0 statico isolatamente non può essere considerato utile per differenziare i singoli parlanti del nostro esperimento, soprattutto se i dati appartengono a stili e tipologie differenti.

Dal punto di vista della identificazione del parlatore, il valore di F0 come variabile statica risponde a quanto da noi affermato in precedenza rispetto al valore linguistico delle porzioni stazionarie. Il valore statico della F0 delle vocali toniche ha principalmente un valore linguistico, inoltre i 4 parlanti di caratteristiche fisiche e di provenienza geografica simile non producono tutti valori di F0 significativamente diversi. I risultati dell'ANOVA risultano essere significativi solo per il confronto della voce VG vs AM. In tutti gli altri casi il risultato del test è non significativo.

Il grafico seguente costruito con i valori medi della frequenza fondamentale in ogni singola voce mostra, forse meglio della tabella post hoc dell'ANOVA, come esistano tipi diversi di voce (VG e AM) e voci invece molto più simili tra loro (LR e SC) almeno sotto il profilo dei valori di F0.

[27] Le voci messe a confronto sono quelle presenti in tabella descrittive statici.

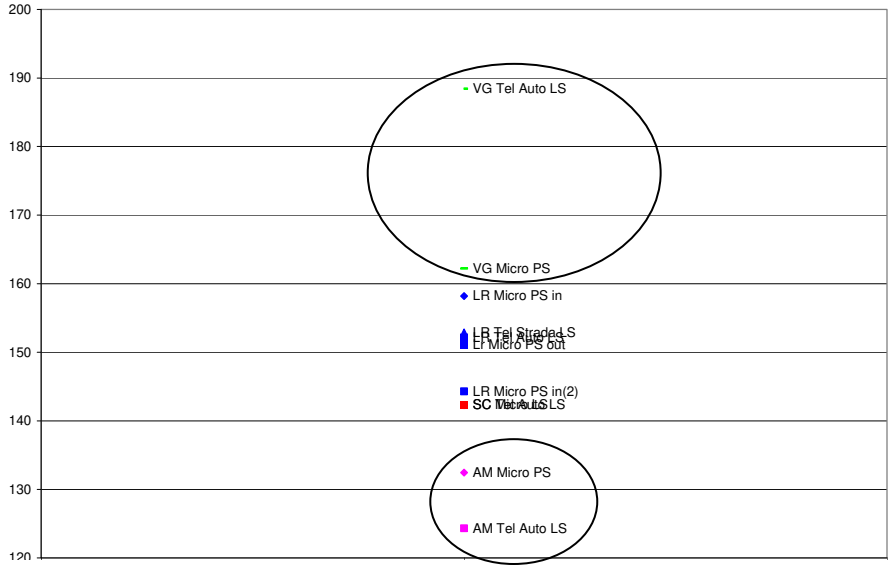


Figura 3: Il grafico presenta i valori medi per ogni parlante in riferimento dello stile e del canale di registrazione

I valori relativi al parlato spontaneo registrati in auto dal parlate LR sono molto più simili a quelli del parlante SC rispetto a quelli registrati dallo stesso parlate LR per telefono o in strada.

10. ANALISI DEI DATI DINAMICI

Comparazioni analoghe a quelle effettuate per i dati statici sono state condotte anche sui dati dinamici (cfr § 4). Di seguito vengono fornite le tabelle riportanti i dati descrittivi per singolo parlante (N=593) e successivamente per ogni singolo canale e stile di voce (N=98). I parlanti scelti per questa comparazione sono: LR, SC e VG.

	N	Media	Dev. std.	Min.	Max.
Parlante LR	593	157,3571	31,56250	77,49	228,68
Parlante SC	593	160,0232	35,62091	79,72	294,30
Parlante VG	593	183,1661	41,77483	74,67	299,35

Tabella 8: Valori medi di F0 dinamico misurato su tutto il segnale, relativi ai singoli parlanti (N=593)

Parlante	LR	SC	VG
Tel. Aula LS	155,6433	148,0813	173,1645
Camera Silente; Voce Alta; LS	188,7603	205,0760	214,3250
Camera Silente; Voce Normale; LS	134,8788	137,8921	174,8402
Amb. Micro LS	160,1450	153,1690	159,0089
Tel. Strada LS	144,0436	157,2374	190,9251
Tel. Auto; LS	161,5613	160,3320	187,8102

Tabella 9: Valori medi di F0 dinamico misurato su tutto il segnale, relativi ai singoli parlanti (N=98) differenziati per canali e stili

Con l'obiettivo di valutare l'effetto su F0 delle variabili parlante e canale-stile si è condotta un'analisi a due vie F0 *by* parlante *by* 'canale-stile'. L'analisi evidenzia un impatto significativo di entrambe le variabili. Infatti, si rifiuta l'ipotesi che mediamente la frequenza fondamentale non differisca tra parlanti. Analogamente si evince che anche il canale-stile contribuisce a differenziare i valori medi di F0. In particolare, invece di verificare il confronto tra i singoli canali e stili, abbiamo testato se l'effetto apportato dal canale avesse un peso maggiore o minore di quello apportato dal parlante. L'*Eta quadrato medio* riporta un valore di 0,230 per il canale e 0,119 per il parlante. Questo rivela che la variabilità intra-parlatore con dati estrapolati in maniera dinamica è maggiore di quella interparlatore.

Sorgente	F	Sig.	Eta quadrato parziale
Modello corretto	50,310	,000	,327
Intercetta	49564,524	,000	,966
Parlante	119,018	,000	**,119**
Canali	105,297	,000	**,230**
Parlante * Canali	8,972	,000	,049

Tabella 10: Test degli effetti fra soggetti (variabile dipendente: frequenza), Indice Eta Quadrato Medio

Un'ulteriore conferma si ricava dall'*effect size* di Cohen (0,29 per il canale e 0,13 per il parlante). In entrambi i casi il valore algebrico maggiore indica il peso maggiore della variabile considerata. La variabile canale apporta maggiore differenza di quanto non faccia la variabile parlante.

Abbiamo quindi mantenuto costante la variabile canale e effettuato la comparazione tra i diversi parlanti. Nella tabella seguente viene presentata la comparazione tra i parlanti LR, SC e VG all'interno della variabile 'Telefono Aula'. Il risultato per questa variabile è confermato: la differenza tra i singoli parlanti è significativa.

Confronti multipli

Variabile dipendente:TelefonoAula

	(I) Parlante	(J) Parlante	Differenza fra medie (I-J)	Errore std.	Sig.	Intervallo di confidenza 95%	
						Limite inferiore	Limite superiore
Tamhane	LR	SC	7,56202*	3,03468	,040	,2528	14,8713
		VG	-17,52121*	3,58314	,000	-26,1560	-8,8864
	SC	LR	-7,56202*	3,03468	,040	-14,8713	-,2528
		VG	-25,08324*	3,52395	,000	-33,5771	-16,5894
	VG	LR	17,52121*	3,58314	,000	8,8864	26,1560
		SC	25,08324*	3,52395	,000	16,5894	33,5771

Tabella 11: Confronto dei diversi parlanti all'interno dello stesso canale (telefono aula)

Non viene però confermato anche all'interno degli altri canali come 'microspia' (stile 'lettura') come mostra la tabella seguente. I valori di significatività mostrano che non vi è alcuna differenza significativa tra i singoli parlanti.

Confronti multipli

Variabile dipendente:Microspia Letto

	(I) Parlante	(J) Parlante	Differenza fra medie (I-J)	Errore std.	Sig.	Intervallo di confidenza 95%	
						Limite inferiore	Limite superiore
Tamhane	LR	SC	6,97607	3,95194	,219	-2,5428	16,4950
		VG	1,13613	5,52730	,996	-12,1997	14,4720
	SC	LR	-6,97607	3,95194	,219	-16,4950	2,5428
		VG	-5,83994	5,41031	,630	-18,9004	7,2205
	VG	LR	-1,13613	5,52730	,996	-14,4720	12,1997
		SC	5,83994	5,41031	,630	-7,2205	18,9004

Tabella 12: Confronto dei diversi parlanti all'interno dello stesso canale (microspia lettura)

11. ANALISI DEI DATI DINAMICO-SELETTIVI

I dati presentati in questo § rappresentano, in parte, il nostro concetto di 'progetto fonetico'. Essi evidenziano l'evoluzione dinamica di tutta la produzione, sebbene la selezione porti a considerare solo il valore medio estrapolato in una parte statica del segnale e solo in quelle porzioni vocaliche dove la sonorità risulta essere un dato rilevante. Inoltre è importante ricordare che con questa analisi vi è la possibilità di avere numerosi dati anche in presenza di poco materiale sonoro. Di seguito come negli altri casi esaminati vengono presentate le tabelle riassuntive dei dati descrittivi per i singoli parlanti e le singole voci (canali e stili), nonché per i singoli segmenti fonici (le vocali).[28]

[28] In questo caso abbiamo aggiunto una variabile, la differenza tra le vocali, per sottolineare l'aspetto selettivo della analisi.

	Media	N	Dev. std.	Mediana	Minimo	Massimo
LR micro Letto	162,0996	231	42,54	154	76,00	342,00
SC Telefono Auto Letto	142,9912	226	24,19	140	97,00	229,00
VG Telefono Auto Letto	180,9249	253	31,47	177	120,00	302,00
VG Micro Parlato Spontaneo	178,8148	135	57,03	171	88,00	348,00
LR Telefono Strada Letto	151,4731	260	32,25	149	81,00	323,00
LR Telefono Auto Letto	154,7761	259	26,99	153	94,00	249,00

Tabella 13: Tabella riassuntiva dei dati descrittivi
per i singoli parlanti e le singole voci

	Vocale /a/	Vocale /e/	Vocale /i/	Vocale /o/
LR Micro Letto	150,11	158,27	174,55	153,38
SC Tel Auto Letto	131,40	158,50	140,30	140,84
SC Micro Letto	132,94	156,66	147,62	140,11
VG Tel Auto Letto	173,33	200,60	200,80	182,00
LR Micro PS	155,20	161,33	150,33	164,66
AM Micro PS	136,00	118,00	147,00	133,12
VG Micro PS	166,50	147,00	171,25	163,33
LR Micro PS	152,40	134,20	156,60	134,00
LR Tel Strada Letto	138,00	184,40	155,80	133,40
LR Micro PS out	153,0000	154,40	135,66	156,25
AM Tel Auto Letto	124,0000	133,40	121,75	117,60
LR Tel Auto Letto	141,4000	178,00	165,00	124,20

Tabella 14: Valori medi di F0 statico misurato solo sulle vocali toniche
diviso per vocale, canale e parlante

Concentrando l'attenzione sulla differenza tra le vocali, abbiamo confrontato i dati estrapolati dalle vocali toniche con quelli estrapolati dalle vocali atone ed inoltre i dati estrapolati dalle vocali definite *good* (cfr. Legenda) e quelle definite *bad*. Ovviamente la qualità vocalica non entra in gioco nelle misure della frequenza fondamentale (o almeno non è così rilevante come la struttura formantica).

Le differenze risultano essere significative nel caso del confronto dei singoli parlanti, mentre il confronto delle singole vocali o dei tipi *bad* e *good* non risulta essere statisticamente significativo.

Vocale	Media	N	Dev. std.	Mediana	Minimo	Massimo	Varianza
VTG	160,12	74	36,217	153,50	101	335	1311,697
VTB	168,35	23	58,255	147,00	96	333	3393,601
VAG	160,79	80	36,237	157,50	80	342	1313,131
VAB	162,05	44	54,759	146,00	76	339	2998,603
Totale	161,60	221	42,872	153,00	76	342	1838,004

Tabella 15: Parlante LR valori di F0 misurato su tutte le vocali presenti differenziate per tipologia[29]

I dati dinamico-selettivi al momento hanno mostrato una buona rappresentazione dei singoli parlanti anche con dati estrapolati da registrazioni su canali differenti e con stili diversi. Gli stessi dati, come già detto, rispecchiano più degli altri la nostra idea di 'progetto fonetico', presentando il percorso che porta al raggiungimento del *target* più che il target stesso. La variabilità inter-parlante risulta essere maggiore di quella intra-parlante.

12. ANALISI DELLA *ARTICULATION RATE*

La velocità è sicuramente un parametro importante nell'eloquio e nel riconoscimento ed identificazione di un parlante. Al momento viene utilizzato solo congiuntamente con altri parametri come la frequenza fondamentale e le frequenze formantiche. L'importanza dell'utilizzo di tale parametro risiede nel fatto che né il canale, né il rumore di sottofondo, dovrebbero influire sulla velocità e sulla estrapolazione dei dati numerici. Ciò che potrebbe influire è invece lo stile (lettura rispetto a parlato spontaneo) e la variabile diafasica (il contesto situazionale). Di seguito sono riportati i valori descrittivi dei singoli parlanti, dei singoli canali divisi per parlante e infine degli stili divisi per parlante.

Per SC e VG non disponiamo di tutti i dati come si può facilmente notare nella tabella 13.

Parlante	Media	N	Dev std.	Mediana	Minimo	Massimo	Varianza
LR	6,52	192	,81113	6,58	3,50	8,04	,658
VG	7,33	192	,80045	7,38	3,47	12,00	,641
SC	6,27	192	1,02779	6,13	4,29	9,53	1,056

Tabella 16: Misure descrittive sull'*Articulation Rate* differenziati per parlante

[29] Le misure sono state effettuate con *Praat* su tutto il materiale registrato.

	LR	SC	VG
Canale	Media	Media	Media
Telefono Aula	7,2758	7,1100	7,3254
Camera silente voce A	6,0371		
Camera silente voce N	7,2792	6,4796	7,4992
Camera silente voce B	6,1608		
Microspia Letto	6,3558	5,5862	7,5996
Microspia Spontaneo in	6,0537	7,5533	7,3175
Microspia Spontaneo out	6,6825		
Lezione	6,4054		
Telefono Strada	6,7517	6,4096	7,5146
Telefono Auto	6,2425	5,6792	7,5713

Tabella 17: Valori medi di AR differenziati per parlante, per modalità e per canale

	LR		SC		VG	
Stile	Media	N	Media	N	Media	N
frasi lette	6,7810	120	6,2587	121	7,5020	120
Parlato Spontaneo	6,3681	48	7,5796	23	7,3175	24

Tabella 18: Valori medi di AR differenziati per parlante e per stile

	Somma dei quadrati	df	Media dei quadrati	F	Sig.
Fra gruppi	119,192	2	59,596	75,919	**,000**
Entro gruppi	449,805	573	,785		
Totale	568,997	575			

Tabella 19: Analisi ANOVA tra i diversi parlanti (variabile AR)

Le differenze tra i parlanti senza considerare i diversi canali e stili sono statisticamente significative.

Nei confronti multipli però, la significatività persiste esclusivamente tra LR e VG e tra VG e SC, negli altri casi invece la differenza non è significativa. Sicuramente LR e VG hanno due velocità molto differenti mentre non è lo stesso per LR e SC.

	(I) Parlante	(J) Parlante	Differenza fra medie (I-J)	Errore std.	Sig.
Tamhane	LR	VG	-,81505*	,08224	**,000**
		SC	,25047*	,09449	**,025**
	VG	LR	,81505*	,08224	**,000**
		SC	1,06552*	,09402	**,000**
	SC	LR	-,25047*	,09449	**,025**
		VG	-1,06552*	,09402	**,000**

Tabella 20: ANOVA tra i parlanti senza considerare gli stili e i canali
(variabile dipendente: AR – la differenza media è significativa al livello 0.05)

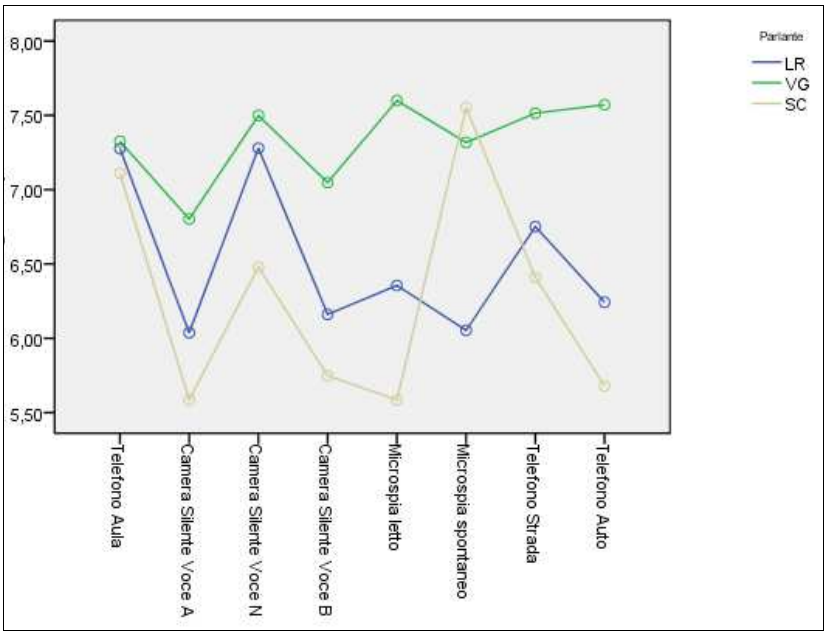


Figura 4: La figura mostra i valori medi per parlante, per canale e per stile

Osservando il grafico delle medie per singola voce e singolo parlante, si nota come sia presente una velocità differente tra i parlanti VG e LR. SC invece in parte segue l'andamento di LR almeno per gli stili 'Telefono Aula', e 'Camera Silente' (Voce A, N e B), mentre registra al suo interno, il valore più basso nella lettura ('Microspia Lettura') e il più alto valore nel parlato spontaneo ('Microspia Parlato Spontaneo'). Tale dato risalta e negli altri due casi LR e VG ha valori contrari. Certo molte sperequazioni potrebbero essere fatte

sulle caratteristiche dei parlanti, sulla loro competenza e abitudine alla lettura o anche sulla gestione del parlato spontaneo e della loro fluenza.

Un ulteriore dato riguarda lo stile voce Alta, Normale e Bassa. In questo caso visto che il parametro non interessa le frequenze, anche la voce con modalità bassa è stata considerata. Come si nota in tutti e tre i parlanti la modalità normale raggiunge velocità maggiori, molto differenti rispetto alla velocità Alta o Bassa. Rammentiamo che il materiale prodotto è identico così come il canale e la modalità di registrazione.

Anche in questo caso abbiamo misurato il peso della singole variabili: parlante e canale. L'Eta quadrato medio rivela, con il suo valore di 0,287 rispetto a 0,208, la maggiore (anche se relativa) importanza del parlante rispetto al canale.

Anche nel confronto tra lo stile 'Lettura' e lo stile 'Parlato Spontaneo' (per il parlante LR è presente anche una registrazione in contesto differente: 'Lezione in Aula') le differenze non sono statisticamente significative.

13. CONCLUSIONE

L'identificazione è il risultato secondario di un processo di discriminazione di una voce. Se due entità devono essere discriminate attraverso i loro attributi allora queste, se differenti, devono differire nei loro attributi. Così, se due persone vengono discriminate e riconosciute attraverso la loro voce allora devono differenziarsi ed essere riconosciute attraverso la loro voce.

La voce, così intesa, è un oggetto multidimensionale e come tale deve essere trattato. Riteniamo che solo la competenza di un esperto possa aiutare a scegliere la dimensione più adeguata e la composizione delle differenti dimensioni. Non tutte le caratteristiche, infatti, aggiungono informazione al processo di comparazione, e non tutte le caratteristiche hanno lo stesso peso (statistico) e lo stesso carico informativo.

Questo lavoro non ha le pretese di modificare le condizioni generali delle comparazioni foniche (SR) ma solo di verificare sperimentalmente il peso di ogni singola variabile e soprattutto di valutare la variabilità inter e intraparlante in funzione degli stili di parlato e dei canali di registrazione.

I risultati ottenuti in questo lavoro si differenziano in base ai parametri e alle variabili considerate.

Considerando le variabili definite statiche, i risultati ottenuti dimostrano che la modalità della voce influenza consistentemente i valori della frequenza fondamentale (parametro considerato molto importante nelle comparazioni foniche). Voce Alta, Normale o Bassa producono dati acustici molto differenti (statisticamente rilevanti e significativi) tra loro anche all'interno dello stesso parlante.

Inoltre anche il canale e lo stile influenzano molto la produzione tanto da far sì che lo stile 'letto' di un parlante venga confuso con il parlato 'spontaneo' di un parlante differente. Nei casi in cui le voci considerate e da comparare, sono molto differenti tra di loro, allora la differenza risulterà statisticamente significativa anche se i dati provengono da stili differenti o da registrazioni attraverso canali differenti. In tutti gli altri casi (cioè con voci mediamente simili) il rischio di falsa identificazione è troppo alto. Spesso in tutti i dati da noi analizzati sembra essere molto più influente il canale di registrazione che il parlante.

La variabilità interna della frequenza fondamentale sembra essere maggiore rispetto a quella misurata tra i differenti parlanti. Qualora si ritenesse utile utilizzare le variabili statiche sarà necessario considerare gli stessi stili di produzione e le stesse modalità.

Per quanto riguarda invece i dati dinamici, i risultati ottenuti non sono affatto confortanti. Le voci vengono spesso confuse e i canali di registrazione risultano avere un carico informativo maggiore dello stesso parlante.

I dati dinamico-selettivi al contrario, al momento hanno mostrato una buona rappresentazione dei singoli parlanti anche con dati estrapolati da registrazioni su canali differenti e con stili diversi. Tali dati come già rappresentato in precedenza, rispecchiano maggiormente la nostra idea di 'progetto fonetico', presentando di fatto il percorso che porta al raggiungimento di un target più che il target stesso. In aggiunta, in questo caso non è stata riscontrata differenza significativa neppure all'interno della diversa qualità vocalica.

Riconsiderando le nostre ipotesi di partenza possiamo concludere che i dati dinamico-selettivi mostrano una minore variabilità intra-parlante ed una maggiore variabilità inter-parlante. Non possiamo affermare lo stesso né per i dati statici né per quelli dinamici.

Riguardando le caratteristiche che nei paragrafi precedenti sono state esposte nei confronti delle variabili, possiamo concludere con la tabella seguente:

Caratteristiche Variabile	**Risultati parziali**
mostra una alta variabilità **inter**parlatore e una bassa variabilità **intra**parlatore	questo è vero esclusivamente per i valori di F0 del parametro Dinamico-Selettivo
è resistente al camuffamento	è stata evidenziata una grande differenza tra Voce Alta e Voce Bassa sia per F0 che per l'AR. Per quanto riguarda i canali invece questi non sembrano camuffare molto i parametri AR e Dinamico-Selettivi
ha una alta frequenza di occorrenza	questo è vero solo nel caso dei parametri dinamici.
è robusta durante la trasmissione	l'unico parametro che sembra non essere influenzato dal canale è l'AR.
è relativamente facile da identificare e misurare	sia F0 (dinamico) che la durata sono molto facili da identificare e misurare.

Ovviamente questo non è che il primo lavoro basato solo ed esclusivamente sulla frequenza fondamentale e sull'*Articulation rate*. Nel prossimo futuro ci concentreremo anche sulle frequenze formantiche e sulla qualità vocalica.

14. BIBLIOGRAFIA

Aitken, C.G.G. (1995), *Statistics and evaluation of evidence for forensic scientist*, Chichester: Wiley.

Barlow, M. & Wagner, M. (1998), Measuring the dynamic encoding of speaker identity and dialect in prosodic parameters, in *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Nov 30 – Dec 4, 1998 (R.H. Mannell & J. Robert-Ribes, editors), Sydney: Australian Speech Science and Technology Association, 81-84.

Basile, C. & Lana, M. (2009), L'attribuzione di testi con metodi quantitative: riconoscimento di testi gramsciani, in *Atti del Convegno Nazionale Ass.I.Term. Terminologia analisi testuale e documentazione nella città digitale*, Aida Informazioni: Roma.

Clermont, F. & Itahashi, S. (1999), Monophthongal and diphthongal evidence of isomorphism between formant and cepstral spaces, in *Proceedings of the Spring Meeting of the Acoustical Society of Japan*, Meiji University Press, 2005-6.

Fant, G. (1960), *Acoustic Theory of speech production*, Hague: Mouton.

Frye (1923), Frye vs. United States (1923), *293 Federal Reports (1st series)*, 1013, 1014 (CA).

George, D. & Mallery, P (2003), ED425201, *SPSS for Windows Step by Step: A Simple Guide and Reference 11.0 update*, Allyn & Bacon, A Viacom Company: Needham Heights, MA.

Gruber, J.S. & Poza, F.T. (1995), Voicegram identification evidence, *American Jurisprudence Trials*, 54, Lawyers Cooperative Publishing.

Ezzaidi, H., Rouat, J. & O'Shaughnessy, D. (2001), Towards combining pitch and MFCC for speaker identification systems, in *Proceedings of Eurospeech 2001*, Aalborg, Denmark, September 3-7.

Hollien, H. (1990), *The acoustics of crime*, New York: Plenum.

Kersta, L.G. (1962), Voiceprint identification, *Nature*, 196, 1253-7.

Künzel, H.J. (1994), Current approaches to forensic speaker recognition, in *Proceedings ESCA Workshop on Automatic Speaker Recognition Identification Verification*, Martigny, Switzerland, April 1994, 135-41.

Künzel, H.J. (1997), Some general phonetic and forensic aspects of speaking tempo, *Forensic Linguistics*, 4, 48-83.

Ladefoged, P. (1993), *A Course in Phonetics*, 3rd edition, Sydney: Harcourt Brace College Publishers.

Laver, J. (1994), *Principles of Phonetics*, Cambridge:Cambridge University Press.

Lindbloom, B. (1990a), Explaining phonetic variation: a sketch of the H&H Theory, in *Speech Production and Speech Modelling* (W.J. Hardcastle & A. Marchal, editors), Dordrecht: Kluwer Academic Publishers, 135-152.

Lindbloom, B. (1990b), On the notion of possible speech sound, *Journal of Phonetics*, 18, 135.

McCormk, P. & Russell, A. (editors) (1996), *Proceedings of the 6th Australian International Conference on Speech Science and Technology*, Canberra: ASSTA.

McDermott, M.C., Owen, T. & McDermott, F.M. (1996), *Voice Identification: The Aural Spectrografic Method*, Colonia, NJ: Owl Investigations Inc.
http://tapeexpert.com/pdf/voiceidauralspectro.pdf

McDougall, K. (2006), Dynamic features of speech and the characterization of speaker: Toward a new approach using formant frequencies, *The International Journal of Speech, Language and the Law*, 13, 89-126.

Nakasone, H. & Beck, S.D. (2001), Forensic automatic Speaker recognition, in *Proceedings of the 2001 Odyssey Speaker and Language Recognition Workshop*, 1-6, Crete, Greece, June 18-22.

Napoletani, D., Romito, L., Sauer, T. & Struppa, D. (in stampa), *Functional dissipation for speaker recognition.*

Napoletani, D., Sauer, T. & Struppa, D. (2002) Patent title: *Functional Dissipation Classification of Retinal Images*.

Nolan, F. (1997), Speaker recognition and forensic phonetics, in *The Handbook of Phonetic Sciences* (W.J. Hardcastle & J. Laver, editors), Cambridge: Cambridge University Press, 744-67.

Osanai, T., Tanimoto, M., Kido, H. & Suzuki, T. (1995), Text-Dependent Speaker Verification using Isolated Word Utterances based on Dynamic Programming, *Reports of the National Research Institut of Police Science*, 48, 15-19.

Romito, L. & Galatà, V. (2008a), Speaker Recognition in Italy: evaluation of methods used in forensic cases, in *Actas del IV Congreso de Fonética Experimental* (A. Pamies & E. Melguizo, editors), Granada, Spagna, 11-14 febbraio 2008 (= *Language Design*, Special Issue, vol. 1), Método Ediciones: Granada, 229-240.

Romito, L. & Galatà, V. (2008b), *Primula: un corpus ristretto di voci calabresi per la valutazione delle metodologie e dei sistemi di riconoscimento del parlatore*, Università della Calabria.

Romito, L., Maddalon, M. & Trumper, J. (1996a), Atteggiamento della Magistratura nei confronti delle perizie foniche, in *Caratterizzazione del parlatore* (F. Fedi & A. Paoloni, editors), Atti delle VI Giornate di Studio del Gruppo di Fonetica Sperimentale (Roma, 23-24 novembre 1995), Roma: Fondazione Ugo Bordoni, 34-45.

Romito, L., Maddalon, M. & Trumper, J. (1996b), La parametrizzazione nei test di riconoscimento, in *Caratterizzazione del parlatore* (F. Fedi & A. Paoloni, editors), Atti delle VI Giornate di Studio del Gruppo di Fonetica Sperimentale (Roma, 23-24 novembre 1995), Roma: Fondazione Ugo Bordoni, 87-93.

Romito, L. (2009), *'Le intercettazioni', contributo a 'Ndrangheta: L'educazione e le istituzioni per un progetto comune*, Università della Calabria, 6 marzo 2009.

Romito, L., Galatà, V. & Lio, R. (2006), Fluency Articulation and Speech Rate as new parameters in the Speaker Recognition, in *Actas del III Congreso de Fonética Experimental*, Santiago de Compostela, 26-24 ottobre 2005, Santiago de Compostela: Xunta de Galicia, 537-549.

Romito, L. & Lio, R. (2008), Stabilità dei parametri nello *Speaker Recognition*: la variabilità intra e interparlatore, in *La Fonetica Sperimentale: Metodo e Applicazioni*, Atti del 4° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Arcavacata di Rende (CS), 3-5 dicembre 2007 (L. Romito, V. Galatà, R. Lio, editors), Torriana: EDK Editore, 125-128 (abstract).

Romito, L., Tucci, M., & Cavarretta, G. (2009), Verso un formato standard nelle intercettazioni: archiviazione, conservazione, consultazione e validità giuridica della registrazione sonora, *Atti del convegno Convegno Internazionale Ass.I.Term*, Università della Calabria, 6-8 giugno 2008.

Romito, L., Turano, T., Loporcaro, M. & Mendicino, A. (1997), Micro- e macrofenomeni di centralizzazione vocalica nella variazione diafasica: rilevanza dei dati fonetico-acustici per il quadro dialettologico del calabrese, in *Fonetica e fonologia degli stili dell'italiano parlato* (F. Cutugno, editor), Atti delle VII Giornate di Studio del Gruppo di Fonetica Sperimentale, Napoli, 14-15 novembre 1996, Roma: Esagrafica, 157-175.

Rose, P. (2002), *Forensic Speaker identification*, London: Taylor & Francis.

Tagliavini, C. (1982), *Le origini delle lingue neolatine*, Patron: Bologna.

Tosi, O. (1979), *Voice Identification: Theory and Legal Applications*, Baltimore: University Park Press.

Trumper, J.B. (1979), *Sociolinguistica giudiziaria*, Padova: CLESP.

VIAAS, (1991), *Voice Identification and Acoustic Analysis SubCommittee*, della *International Association for Identification*, pubblicato negli Atti dell'Associazione VCS.

Zavattaro, D. (2005), *Articulation Rate e l'identificazione del parlatore a scopo forense*, Tesi di di dottorato in Scienze Forensi, XVIII ciclo, A.A. 2004-2005, 2a Università di Roma Tor Vergata.

LOUDNESS E 'LIVELLO DEL DIALOGO' NELLE TRASMISSIONI RADIOTELEVISIVE

Mauro Falcone [a], Antonino Barone [b], Alessandro Bonomi [b],
Alessandro Balestri [b], Anna Grazia Santoro [b], Maria Dell'Osso [b]
[a] Fondazione Ugo Bordoni
[b] Istituto Superiore delle Comunicazioni e delle Tecnologie dell'Informazione
mfalcone@fub.it, antonino.barone@sviluppoeconomico.gov.it

1. SOMMARIO

Il segnale audio che riceviamo attraverso i media (radio, tv, internet, ecc.) può essere, e di fatto lo è, pesantemente, affetto da diversi tipi di elaborazioni e alterazioni. È un fatto ben noto che questi segnali, e quindi anche la voce, sono codificati, e quindi compressi perdendo parte della loro originaria informazione, secondo diversi standard (mpeg2, mpeg4, ecc.). Meno noto è il fatto che questi segnali possono essere elaborati in modo da modificarne, con diversi fini, il loro contenuto energetico. In particolare il segnale è sicuramente manipolato (rispetto ai suoi naturali livelli) in fase di mixing, ma ulteriori e più o meno arbitrarie modifiche sono possibili successivamente. Solo negli ultimi anni si è iniziato a studiare il problema del loudness nelle trasmissioni radiotelevisive cercando di risolvere il problema sia del dislivello 'channel to channel', sia del dislivello 'program to program' e infine quello del 'program to advertising'. Quest'ultimo caso in particolare è stato oggetto di diverse indagini sia in quanto soggetto a normative giuridiche, sia perché aspetto percepito come particolarmente fastidioso dagli ascoltatori. Con la raccomandazione internazionale ITU-R 1770 e le sue successive modifiche si è dato un primo fondamentale contributo alla soluzione del problema, risolvendo che tipo di misura deve essere effettuata per misurare il livello del segnale audio. L'unità di misura secondo la predetta raccomandazione è il 'Loudness Unit' (LU), che è pur sempre una misura in decibel. Per semplicità di lettura nell'articolo riporteremo tutte le misure con l'abbreviazione 'dB' anche le misure di loudness secondo la ITU-R 1770. Definito pertanto 'come' misurare il livello, rimane (almeno) ancora un secondo e fondamentale punto da risolvere, ovvero 'quando' effettuare la misurazione. Non è infatti corretto misurare il livello indiscriminatamente su tutto il segnale audio trasmesso per quantificare correttamente il livello del loudness, ovvero del volume percepito dall'ascoltatore, ma devono essere selezionati solamente quelle parti percettivamente rilevanti, trascurando tutto il rimanente. A tal fine sono oggi utilizzati, nella maggior parte dei casi, due diversi approcci: il 'dialogue intelligence' ed il 'gating'. Nel primo caso il sistema opera una caratterizzazione del segnale in 'parlato' e 'non parlato' per poi eseguire la misurazione solo sul primo tipo (ipotizzando appunto che in ogni caso la veicolazione maggioritaria delle informazioni avviene attraverso la voce), nel secondo caso invece si definisce una soglia di riferimento tale che tutto ciò che ha valore superiore viene considerato di interesse per la misura, mentre tutto ciò che è inferiore viene tralasciato. Ognuno di questi metodi ha specifici vantaggi, ed ovviamente svantaggi. Se il 'dialogue intelligence' può essere operato online senza conoscere il livello medio dell'intensità del segnale, a suo svantaggio c'è il fatto che non può essere utilizzato su segnali musicali, e che risulta computazionalmente complesso. La tecnica del 'gating' al contrario è di facile realizzazione, può essere applicata su qualsiasi tipo di segnale audio.

Non ha infatti senso un gating assoluto senza conoscere quale sia il livello medio di intensità sonora dell'audio che stiamo considerando.

In questo lavoro si mostrano i risultati di un'ampia campagna di misura effettuata attraverso strumentazione professionale e l'utilizzo dello strumento LM100 della Dolby, che è oggi il riferimento internazionale per la misura del loudness attraverso la tecnica di 'dialogue intelligence'. Per quanto riguarda il 'gating', invece, è stato sviluppato un apposito software dagli autori. Sfortunatamente nell'utilizzo del 'dialogue intelligence'. essendo vincolati al funzionamento dello strumento utilizzato, non è possibile variare alcuna funzionalità in quanto il software utilizzato per il 'voice detection' è in ogni caso protetto da brevetto. Al contrario il software sviluppato permette una facile configurazione dei parametri e realizza sia la misura di intensità in RMS, sia la normativa ITU-T 1770. La misura di loudness proposta in questa normativa è subito stata accettata con entusiasmo sia dalla comunità scientifica, sia dalle industrie legate alla produzione e diffusione dell'audio nel broadcast, ed è stata implementata nel nostro sistema di misura. Questa misura tiene conto sia dell'effetto di interferenza dell'ascoltatore, o meglio della sua testa che viene approssimata come una sfera di circa 21 cm, sia di una curva di adattamento relativa alla sensibilità dell'apparato uditivo umano per i suoni della stessa classe di quelli comunemente trasmessi dalle televisioni. Queste curve di 'adattamento' (curve denominate con le lettere: A per la telefonia, B per i segnali di media qualità, e via così per C, D ecc.) sono ben note in psicoacustica, e nella raccomandazione in questione, in particolare, si propone una revisione della curva B.

Le due tecniche, 'dialogue intelligence' e 'gating', vengono messe a confronto su un'ampia ed eterogenea quantità di programmi, relativi a tutte le principali emittenti nazionali. Da questa prima analisi si evince come sia possibile trovare un'equivalenza tra le due metodologie solo parzialmente, e sotto vincoli molto stringenti sia della tipologia dei contenuti e sia della corretta, o perlomeno omogenea, realizzazione del materiale nella fase di missaggio. Vengono riportati i risultati di campagne di misura atte a studiare e quantificare tutte e tre le situazioni di confronto del loudness (C2C, P2P, P2A). Il materiale utilizzato è stato acquisito nell'anno corrente principalmente dalle emittenti RAI e Mediaset, e nell'orario di prima serata e comunque sempre nell'arco di maggior ascolto. Tutto il materiale audio è manualmente etichettato a due livelli: un primo livello individua inizio e fine programma secondo i palinsesti relativi, e un secondo livello dove il segnale viene diviso in due classi ovvero A (il programma vero e proprio) ed R (che contiene tutto il resto come spot, promozioni, jingle, prossimamente, ecc.).

I risultati sperimentali riportati costituiscono un punto di riferimento importante, in quanto non ci risulta siano disponibili pubblicamente studi simili a questo. Inoltre lo studio tra le diverse tecniche di selezione è, anch'esso, una novità nel panorama degli studi sul loudness, e vuole essere un primo contributo alla soluzione di questo problema che già vede schierati i diversi produttori di strumentazione di misura su due fronti opposti. Infine le investigazioni pilota sulle differenze dello stesso contenuto audio attraverso diversi media, e lo studio delle possibili alterazioni dei livelli di parlato rispetto ad un'acquisizione lineare di laboratorio costituiscono i punti di partenza per lo sviluppo di nuove attività sperimentali di ricerca.

2. LOUDNESS: OVVERO IL LIVELLO SONORO PERCEPITO

2.1 Da una misura fisica ad una misura informatica (passando per una elettronica)

Un segnale audio è sempre associato ad una sua "intensità" sonora, ovvero a quello che più comunemente chiamiamo 'volume', o a volte più specificatamente 'potenza' del segnale sonoro. Insomma quella grandezza che ci indica se un segnale audio è più o meno forte. Poiché qualsiasi suono esiste solo se esistono vibrazioni in un mezzo (tipicamente l'aria) e queste vibrazioni sono misurabili, o percepibili, grazie alla variazione di pressione che provocano ne discende che la misura di potenza sonora sarà proporzionale alla pressione associata al segnale audio. Ricordando che la pressione si misura in pascal, dove un pascal, equivalente ad una forza pari ad un newton per metro quadro. Ricordando infine che all'incirca un peso di 102 grammi poggiato su un tavolo opera, sul nostro pianeta, una forza di circa un newton allora possiamo avere un'intuitiva idea della pressione pari ad un pascal pensando ad un pannello largo un metro quadrato e del peso di circa 102 grammi. Per comodità tuttavia la potenza sonora si misura in scala logaritmica e in decibel dove la formula che lega decibel e pressione è tale che lo zero dB SLP (Sound Pressure Level) equivale ad una pressione di 20 micro pascal, detta anche p0. Quindi la misura fisica del volume di un suono e il suo valore SPL dB e si esprime in decibel per convenienza. Questa è una misura assoluta, cioè non ambigua e primaria, e la si può ottenere solo con sistemi di misura professionali e ben tarati quali i fonometri. Esistono in realtà tutta una serie di misure possibili dell'intensità o del volume di un suono (di picco, di valor medio, ecc.), l'importante è avere bene in mente che comunque finché parliamo di misure fisiche del segnale audio facciamo sempre riferimento a misure di pressione del tipo

$$SPL_{dB} = 20 * \log_{10}\left(\frac{p}{p_0}\right) \tag{1}$$

e che queste sono direttamente legate alla nostra percezione dei suoni, ovvero al fenomeno acustico vero e proprio. Ad esempio un normale dialogo tra persone avviene a circa 74 dB SPL, il che significa che il valore di p (ovvero della pressione) è circa 5000 volte maggiore del valore di riferimento p0, mentre la sirena di un'autoambulanza raggiunge circa un valore di un milione di volte p0 (ovvero 120 dB SPL). Tuttavia non studiamo quasi mai, ad eccezione appunto delle misure di rumore ambientale o similari attività, i segnali audio nel loro originario dominio, ovvero come onde acustiche cioè come variazioni di pressione, ma siamo indotti a realizzarne una copia 'analoga' (tipicamente come segnale elettrico) di più facile rappresentazione, utilizzo e studio. Fino alla rivoluzione del 'digitale' pionieristicamente iniziata circa 20 anni fa e compiuta solo nei nostri giorni gli studi sul segnale audio, e quindi anche sul suo livello e sulla sua potenza, sono stati traslati sul segnale analogico elettrico, anche qui non senza problemi di standard, di riferimenti, ecc. Tralasciamo completamente questa pure importante fase storica, e passiamo direttamente alla odierna rappresentazione del segnale audio in forma digitale ovvero in forma del tutto informatica del segnale che quindi non è più rappresentabile come segnale fisico 'analogo' a quello di interesse ma è piuttosto una sua rappresentazione formale. Un segnale digitale non ha quindi, in linea del tutto teorica, alcuna relazione con la sua potenza se non a livello di variazione che viene ovviamente sempre mantenuta. In altre parole dovremmo definire le

caratteristiche di tutto quel sistema di trasmissione e trasformazione del segnale digitale fino alla sua ultima realizzazione fisica attraverso i sistemi di diffusione sonora, per sapere quanto un certo segnale audio digitale suonerà forte nel momento della sua realizzazione. Questa non semplice operazione è tipicamente definita soltanto nell'ambito professionale (ad esempio negli studi di registrazione), a volte è operata per ottenere garanzie di qualità conformi a ben definiti standard (come ad esempio nelle sale cinematografiche), ma non è mai onorata in ambito consumer. In ambito consumer non esiste quindi una procedura per cui partendo dalla nostra sorgente audio digitale (CD, DVD, DTT, ecc.) l'utente abbia la possibilità di riprodurre il segnale audio con un livello pari a quello originariamente definito per una sua 'ideale' fruizione. Non stiamo ovviamente proponendo che un determinato livello di ascolto sia imposto all'utente consumer (come per altro avviene nelle sale cinematografiche), ma che l'utente abbia un riferimento corrispondente al livello originario pensato dal chi ha realizzato quel segnale. Se ad esempio in un film il normale dialogo è stato realizzato per essere proposto all'utente ad un determinato volume, in maniera tale che poi un successivo segnale bisbigliato sia appena percettibile, il dialogo dovrà essere fruito a quel livello e non a uno minore, pena la perdita di intelligibilità dei segnali bisbigliati. È necessario insomma non solo produrre tutti i segnali secondo ben condivisi protocolli, ma anche definire una procedura di omologazione della catena di trasmissione e di riproduzione dell'audio. Oggi, purtroppo, entrambe queste fondamentali condizioni sono ampiamente disattese.

2.2 Il loudness nelle comunicazioni radiotelevisive

Particolarmente complesso sembra in campo delle comunicazioni radiotelevisive. Da un lato infatti la radio sta divenendo sempre più un sistema di ascolto fruito da utenti in mobilità (automobile, ecc.), ma al contempo non abbandona il suo ruolo altamente culturale attraverso la trasmissione di musica d'arte, diretta di concerti ed opere ecc. Dall'altro anche la televisione indirizza sullo stesso mezzo di fruizione film e fiction di altissima qualità audio video, e trasmissioni del tipo talk show e reality dove la scarsa qualità del segnale sembra essere un'elemento caratterizzante della trasmissione stessa piuttosto che una necessità tecnica.

A questo punto è doveroso sottolineare il fatto che l'avere a disposizione nuove tecnologie come il digitale terrestre, il satellitare, la televisione su protocollo internet per non parlare poi dei formati in alta definizione anche attraverso i sistemi di codifica di seconda generazione S2 per il satellitare e T2 per il digitale terrestre, permette si tecnicamente una qualità audio potenzialmente di altissimo livello che nessun utente ha mai potuto sperimentare in passato nell'ambito domestico, ma condizione prioritaria e che la creazione e la diffusione dei contenuti seguano modalità adeguate a queste potenzialità e comunque certamente diverse da quelle utilizzate in passato. Tuttavia degli ostacoli ancora sussistono nonostante le più ampie e migliori possibilità tecnologiche, ostacoli che per assurdo sono proprio originati da questa generale migliore situazione. Come prima cosa il sistema di fruizione utilizzato dagli utenti è ancora necessariamente troppo diverso sia per capacità tecnologica sia per modalità di utilizzo. Per la radio si passa dall'ascolto delle informazioni in automobile all'ascolto di concerti attraverso il proprio impianto hi-fi; per la televisione si va dalla fruizione in uno scenario home theatre dei film di recentissima produzione, a quella di fruizione di informazioni o entertainment in ambito casalingo (cucina, ecc.). Inoltre proprio il fatto di avere un migliore e più potente sistema impone uno stretto controllo della qualità del materiale in 'ingest' (per 'ingest' si intende tutto quel

materiale archiviato e pronto per la messa in onda), nonché l'adeguamento di tutti gli studi di produzione alle normative e ai protocolli di qualità interni. Per il primo problema non vi è, stante l'attuale scenario, una soluzione che possa soddisfare tutti. In realtà la questione potrebbe essere risolta attraverso l'utilizzo di metadati (possibilità già ampiamente prevista nelle comunicazioni audio digitali) e nella realizzazione si dispositivi di ricezioni predisposti a configurarsi opportunamente in funzione dei metadati ricevuti. Purtroppo però gli interessi veicolati nelle nuove piattaforme, ed in particolare nel digitale terrestre, sembrano distratti da tutt'altre funzionalità come l'interattività a scapito di una seria gestione della qualità del segnale audio video. Senza questa soluzione non rimane altro che ricorrere a dei compromessi, che comunque non risolvono la situazione, creando canali specifici e tematici ottimizzati alla fruizione in ben definite condizioni e con ben definiti sistemi. Allora è chiaro che un canale 'pay per view' che trasmetta musica classica o film di recente produzione potrà, anzi dovrà, essere fruito sotto certe condizioni, mentre un canale di informazioni o di telepromozioni sarà correttamente fruito in condizioni probabilmente del tutto incompatibili con le precedenti. La permutazione degli ideali e rispettivi sistemi di utilizzo non solo non garantisce un'ottimale fruizione, ma certamente può portare a un forte degrado della qualità percepita in particolar modo quando si vuole utilizzare un sistema di media bassa qualità riproduttiva per riprodurre un segnale di alta qualità. Questo impone anche un diverso stile di produzione legato non solo ai gusti del'audience di quei programmi e alle caratteristiche del precipuo segnale, ma anche all'ipotetica piattaforma con cui quel segnale verrà riprodotto. Insomma si dipinge un panorama articolato e differenziato che non premia la qualità globale e la possibilità di utilizzo di sistemi di riproduzione adeguati, ma piuttosto penalizza e costringe ad un sottoutilizzo le tecnologia già disponibili.

2.3 Le nuove normative internazionali sul loudness

Una prima significativa conquista in tale campo comunque inizia a farsi strada a livello internazionale e nelle raccomandazioni degli organismi regolatori. Conseguentemente le aziende produttrici di strumentazione stanno adeguando, spesso anzi stanno anticipando, i loro sistemi di punta per quanto riguarda le misure di potenza sonora, ed infine i responsabili di produzione studio stanno, per lo meno nei centri di eccellenza, modificando le linee guida di produzione. Si tratta di un segnale che seppur debole è di grande importanza in quanto apre la strada ad un'effettiva rivoluzione nell'ambito della qualità audio che finalmente ha come termine e di riferimento la qualità percepita dall'utente e non più vincoli legati alle necessità di rappresentare il segnale analogico conformemente a requisiti della strumentazione utilizzata. Questa è la grande rivoluzione che l'impatto del digitale porta nei nuovi scenari. Se prima infatti le caratteristiche di un segnale audio, almeno per quanto riguarda la sua intensità, erano definite pensando a contenere gli istanti di massima intensità o picchi entro certi limiti, oggi questi vincoli non sussistono virtualmente più e l'attenzione può essere spostata su ciò che è realmente importante ovvero sul volume percepito ovvero sul loudness.

È cambiato quindi il termine di riferimento: prima era necessario essere sicuri che le strumentazione analogiche fossero correttamente utilizzate ovvero che i segnali non superassero mai certi livelli di picco che avrebbero prodotto degli artefatti, oggi possiamo invece più correttamente richiedere che il segnale abbia delle caratteristiche ben determinate per quanto riguarda quella che sarà la sua percezione all'utente finale. Ovviamente dei requisiti tecnici sono ancora richiesti, ma grazie alle potenzialità del

digitale vanno in secondo piano e non impattano, se non in minima parte, sui requisiti di produzione e trasmissione del segnale audio. Il fatto essenziale è comunque la coscienza sviluppatasi a fronte del fatto che oggi è realmente possibile veicolare segnali di altissima qualità audio anche attraverso canali quali la televisione o la radio. Coscienza che purtroppo non è dimostrata, se non raramente, dai diretti interessati cioè dai singoli broadcaster ma piuttosto dagli organismi di regolamentazione internazionale quali l'ITU (*International Telecommunication Union*) o le associazioni di settore come l'EBU (*European Broadcasting Union*). Queste infatti hanno da subito riconosciuto le grandi potenzialità già disponibili all'utente finale, e si sono prontamente fatte carico di regolare il problema del loudness, che costituisce in realtà solo il primo passo verso una nuova realtà di fruizione della qualità audio.

2.4 La raccomandazione ITU-R BS.1770 e suoi futuri sviluppi

Nel 2006 il gruppo di lavoro sul *Programme production and quality assessment* dell'ITU-R ha pubblicato la prima versione della raccomandazione BS.1770 dal nome *Algorithms to measure audio programme loudness and true-peak audio level*. Con questo documento si è di fatto ufficialmente aperta una nuova era nell'ambito della normativa sulla qualità audio proprio per quanto detto nel paragrafo precedente. Il documento in questione, nella sua stesura originale, definisce un nuovo tipo di pesatura, simile alla già nota curva B già utilizzata in altri campi audiometrici, e tiene conto nella misura del loudness delle interferenze tra onda acustica e il soggetto ascoltatore. Le misure di loudness vengono inoltre definite per i diversi tipi di riproduzione: mono stereo e multicanale. Descriviamo brevemente le caratteristiche salienti di questa originaria raccomandazione anche se, come vedremo in seguito, alcuni suoi aspetti sono già in fase di revisione e di miglioramento.

Il segnale audio viene quindi filtrato con un filtro HS (Head Simulation) che tiene conto delle interferenze dell'ascoltatore con il segnale acustico. In maniera approssimata possiamo assimilare la testa dell'ascoltatore con una sfera di circa 21 cm di diametro. Ne risulta un innalzamento della alte frequenze, superiori a 1 kHz, secondo la funzione di trasferimento illustrata nella figura 1.

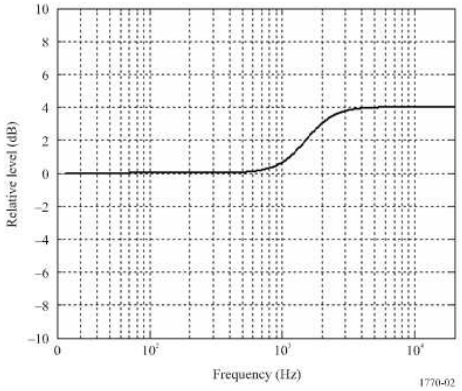


Figura 1: Funzione di trasferimento del filtro HS della 1770 che tiene conto della interferenza della testa dell'ascoltatore

Tale filtro è implementato semplicemente con un filtro numerico del secondo ordine come illustrato nella figura 2. I coefficienti, per un segnale a 48 kHz, sono forniti nella raccomandazione.

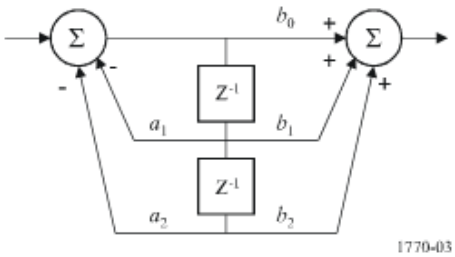


Figura 2: Schema del filtro numerico della 1770

Detto x il segnale di ingesso al filtro, e y il segnale di uscita, e indicando con k i valori interni, il filtro in questione può essere formulato nel modo seguente:

(2) $$y_n = k_n * b_0 + k_{n-1} * b_1 + k_{n-2} * b_2$$ dove

(3) $$k_n = x_n - k_{n-1} * a_1 - k_{n-2} * a_2$$

Anche il secondo filtro, denominato RBL (revised B) è costituito come in figura 2. La sua risposta in frequenza è invece riportata in figura 3, e può considerarsi una rivisitazione della nota curva B adattata allo specifico scenario delle trasmissioni radiotelevisive.

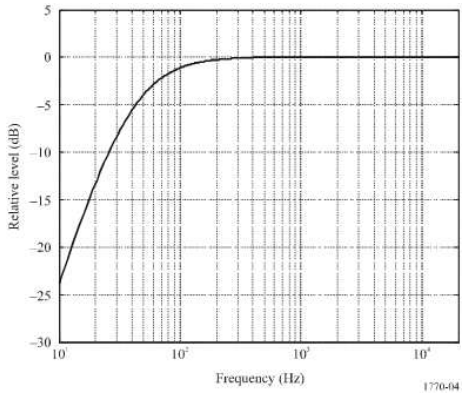


Figura 3: Funzione di trasferimento del filtro RLB della 1770
che tiene conto di fattori psicoacustici di percezione

Come abbiamo detto questa elaborazione è valida per tutti i tipi di segnale: monofonico, stereofonico e multicanale. Tuttavia particolare attenzione va dedicata al complicato multi-canale, ed in generale lo schema da seguire è quello riportato in figura 4, dove i pesi G per i primi tre canali (destro, sinistro e centrale) sino uguali ad uno, mentre gli ultimi due (destro e sinistro posteriori) sono pari a circa 1.5 dB.

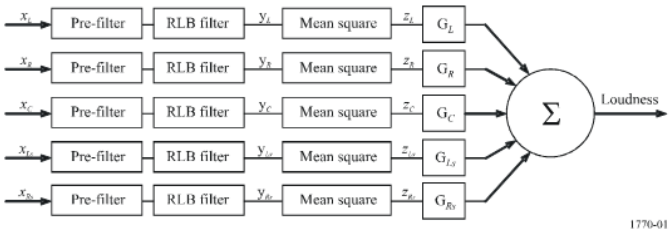


Figura 4: Schema generale per il calcolo del loudness nella 1770

Questo schema generale è oggi già oggetto di revisione ed è ormai certo che a questo vengano aggiunte almeno due nuove parti. Una prima che tenga conto anche del contributo del LFE (*Low Frequency Effect*), ovvero dei cosiddetti subwoofer dei sistemi multicanali 5.1 di cui appunto i cinque canali sono quelli della figura 4, mentre il '1' è rappresentato dal LFE che in ogni caso contribuisce alla percezione del loudness. Il secondo aspetto è invece molto più complesso, e si basa sul principio che tipicamente una persona si crea un'idea della potenza sonora di un segnale non sulla media di tutto il segnale, ma solamente sulla media dei segnali a livello più alto o comunque dei segnali che veicolano un'informazione importante (ad esempio il parlato). Per selezionare solamente il segnale significativo abbiamo quindi due possibilità: individuare tutto il segnale con un'intensità sonora rilevante (ovvero scartare tutto il segnale che potremmo definire di sottofondo), oppure individuare tutto il segnale che convoglia informazioni significative ed in particolare il segnale vocale. In quest'ultimo caso insomma bisogna aggiungere un dispositivo di *Voice Activity Detection* (VAD) che vada a eliminare tutto ciò che non è segnale vocale. Questa è la soluzione ad esempio utilizzata da Dolby e comunemente nota con il nome di *Intelligent Dialogue*. Si noti che tale misura costituisce un fattore essenziale in alcuni standard commercialmente utilizzati nella diffusione di programmi e di proprietà Dolby, ma universalmente utilizzati sia in ambito cinematografico sia in ambito di trasmissioni satellitari dove si fa utilizzo di metadati (il ricevitore satellitare infatti a differenza di quello della televisione digitale terrestre è in grado di utilizzare alcuni metadati tra cui appunto quello del livello del dialogo). Tuttavia l'altro sistema, quello del gating, è di più facile implementazione, può essere applicato su tutti i tipi di segnali, anche dove non c'è voce, e offre risultati altrettanto buoni. Per questi motivi gli organismi internazionali succitati stanno definendo una nuova normativa che tenga conto del gating e del contributo del LFE per il multicanale. Ovviamente l'importanza del gating è di gran lunga di maggiore rilevanza in quanto impatta su tutti i segnali, dal monofonico al multicanale e comunque definisce nuove regole di misura su tutta la catena dalla produzione alla fruizione. Se si vuole una misura che tenga conto della percezione dell'ascoltatore il gating è essenziale, senza questo le misure verrebbero (come è stato fatto in passato) completamente falsate. Poiché il gating va ad eliminare quei segmenti di segnale al di sotto di una certa soglia, il risultato di una misura con il gating sarà sempre maggiore di una misura operata senza il gating. Il loudness misurato secondo la normativa in questione utilizza una scala di intensità sonora che indicheremo come LU, la misura del loudness su un brano o intero programma, cioè il loudness medio di segmento audio è indicato con il termine LKFS (*Loudness K-weighted Full Scal*e).

3. IL DISALLINEAMENTO AUDIO NELLA RADIOTELEVISIONE

Nel paragrafo precedente abbiamo brevemente introdotto il problema di come misurare correttamente il livello sonoro da un punto di vista percettivo nei diversi scenari di comunicazione radiotelevisiva (ma quanto detto è in realtà valido in un generico scenario di fruizione audio). La soluzione a questo problema è importante in quanto propedeutica a realizzare, trasmettere e quindi fruire un audio consistente da un punto di vista della sua intensità sonoro. Ovviamente non vogliamo qui discutere della corretta realizzazione di un segnale audio, di un parlato, ecc., e quindi di problemi legati a registrazioni dal vivo, piuttosto che alle registrazioni in fase di doppiaggio, al successivo missaggio e quanto altro. Diamo per scontato (cosa in realtà tutt'altro che vera nella realtà dei fatti, ma che esula dall'interesse di questo lavoro) che il prodotto audio sia realizzato correttamente, secondo le normali fasi di lavorazione descritte in figura 5, o comunque consideriamo come il risultato delle prime due fasi, registrazione o doppiaggio e missaggio, come corrette per definizione. Vi sono a questo punto tutta una serie di fattori che possono influire sulla qualità audio del segnale e quindi del parlato (codifica, banda passante, qualità della rete trasmissiva, ecc.), ma per quanto riguarda il loudness in letteratura si è soliti distinguere tre specifici problemi di disallineamento del livello sonoro Questi sono per lo più generati da un cattivo, o in alcuni casi dalla totale assenza, di controllo di qualità nella produzione e distribuzione dei programmi e dalla mancanza (o dal non rispetto) di normative tecniche a riguardo.

Figura 5: Le tre fasi di lavoro (doppiaggio, missaggio e messa in onda)

Il primo caso di disallineamento audio è dovuto al fatto che ogni emittente sembra trasmettere segnali audio ad un livello medio del tutto autonomo da quello delle altre emittenti. Questo accade sia tra emittenti di broadcaster diversi, ma anche tra i diversi

canali di uno stesso broadcaster. Il disallineamento tra canale e canale (channel to channel o più brevemente C2C), è prevalentemente dovuto ad una scarsa attenzione alla qualità della messa i onda ed è il principale motivo per cui spesso cambiando da un canale all'altro siamo costretti ad modificare il livello audio con il telecomando per portarlo a quello che è il nostro livello di ascolto preferito.

Un secondo e più grave problema è quello del disallineamento tra un programma e l'altro. Qui il discorso si fa più complesso. A meno che non si stia considerando un'emittente monotematica (tipologia che per altro sta mano a mano guadagnando spazio anche nel nostro paese), diversi programmi possono avere caratteristiche audio completamente diverse (si passa da un telegiornale, ad un film d'azione, a un concerto di musica classica, ecc.). È possibile che questi programmi siano stati realizzati a dei livelli audio diversi, ma qualora si sia controllato il loro livello audio, non è detto che si sia utilizzata una misura di 'loudness', e pertanto questi risulteranno percettivamente disallineati tra loro nonostante si sia eseguito un controllo di qualità. Il fenomeno del disallineamento tra programma e programma (program to program o più brevemente P2P), oltre a sommarsi al precedente, è particolarmente fastidioso perché mostra una disomogeneità nell'ambito di una fruizione continua (siamo sempre sullo stesso canale). Al fine di risolvere questo problema esistono due possibili strade. La prima è utilizzare degli strumenti che riadattino, si noti bene modificandolo artatamente, il volume dei programmi al fine di riportarlo entro determinati valori, la seconda è quella di operare un controllo di qualità in fase di produzione del programma e/o eventualmente nella fase di 'ingest' (la fase di 'ingest' è quella in cui il programma viene acquisito da un archivio o da un distributore e viene messo nei server di distribuzione prima della messa in onda e tipicamente anche il momento in cui viene operato un controllo di consistenza e qualità del programma). Purtroppo la prima soluzione sembra, anche per i suoi minori costi, essere preferita, ovviamente a scapito della qualità dell'audio. Tuttavia nonostante questi accorgimenti, ovvero l'utilizzo di strumenti che adattano il volume dei programmi entro determinati valori, il problema è ancora presente. Ne consegue che una diversa soluzione, come si spera venga presto adottata anche su indicazione degli organismi internazionali che stanno lavorando ad una serie di raccomandazioni a favore di questa strategia, che normalizzi secondo principi percettivi il livello audio di ciascun singolo programma, o meglio di ciascun oggetto audio messo in onda, non solo risolverebbe definitivamente il problema, ma lo farebbe anche ottimizzando la qualità dei segnali trasmessi. Abbiamo sottolineato che a dover essere allineato deve essere ciascun oggetto audio piuttosto che il programma, perché in alcuni casi il programma non è ben definito in quanto composto da diversi elementi, e inoltre è spesso interrotto da altri programmi o dalla pubblicità.

Ed eccoci quindi al terzo e più discusso problema, quello del volume della pubblicità. Il volume della pubblicità nelle trasmissioni radiotelevisive è oggetto di norme di legge. La normativa italiana, infatti, così recita: "è fatto divieto alla concessionaria pubblica e ai concessionari privati per la radiodiffusione sonora e televisiva di trasmettere sigle e messaggi pubblicitari con potenza sonora superiore a quella ordinaria dei programmi". Leggi simili sono anche emanate in altri paesi della Comunità Europea. In molti paesi, come ad esempio la Gran Bretagna, la formulazione è leggermente diversa in quanto la legge chiede che il livello sonoro non arrechi disturbo all'ascoltatore. Anche se la normativa italiana sembra piuttosto chiara tecnicamente, dopo più di venti anni non si è raggiunta una metodologia per espletare i necessari controlli che dimostrino il rispetto

piuttosto che l'infrazione della legge. Di fatto quindi non ci risulta che si sia mai perseguito nessuno sulla base di questa legge nonostante sia evidente a molti che questa non è rispettata, o comunque che la situazione è tale da non essere gradita alla utenza. Il dislivello tra programma e pubblicità (program to advertising o più brevemente P2A) è certamente quello più evidente non solo per i valori di disallineamento che si riscontrano, ma anche perché è quello che si presenta più frequentemente nei palinsesti, se non altro per il gran numero di inserzioni pubblicitarie. Tutto questo rende questo particolarmente importante la risoluzione di questo problema, che tecnicamente tuttavia non è molto diverso da quello precedentemente descritto del P2P. Infatti il singolo spot pubblicitario, così come la telepromozione o la sigla o quanto altro, possono considerarsi dei programmi veri e propri anche se di breve durata. Possono considerarsi programmi non solo perché prodotti autonomamente e indipendentemente dal programma che li contengono, ma anche perché comunicano un messaggio diverso e autonomo da quello del programma che li contiene. Quindi quanto detto per il caso precedente del 'program to program' può applicarsi anche per le pubblicità, per le sigle, per le telepromozioni e in generale per tutti quegli oggetti definiti 'interstiziali' cioè di breve durata e inseriti nel contesto omogeneo di un programma. Ciò premesso è necessario considerare che vi è, purtroppo, un fattore che differenzia questo caso dai precedenti. Per la pubblicità, le sigle, ecc. può esservi l'interesse di fare in modo che questi suonino ad un livello percettivamente più forte rispetto agli altri programmi in quanto oggetti sonori che devono 'risvegliare' l'attenzione dell'ascoltatore (la pubblicità deve richiamare l'attenzione, la sigla deve evidenziare l'inizio del programma, e così via). Una misura di loudness ideale dovrebbe rispecchiare fedelmente il livello sonoro percepito, e quanto definito nella raccomandazione ITU-T BS.1770 è sicuramente una buona soluzione per i segnali radiotelevisivi, ciò nonostante è possibile, seppur in misura limitata, realizzare sei segnali che a parità di misura, ovvero che abbiano lo stesso valore di LKFS, suonino percettivamente diversi. Le tecniche di gating o di dialogue intelligence possono ridurre molto questo problema fino a quasi a ridurlo. Descrizioni statistiche di ordine superiore che vanno a considerare la distribuzione degli istogrammi della energia, mostrano che a parità di misura del LKFS, se i segnali da comparare hanno una diversa qualità, possono essere effettivamente percepiti come di diversa intensità. Pertanto, specialmente per la pubblicità e tutti gli interstiziali di breve durata, è necessario, affinché questi vengano percepiti come omogenei rispetto ai programmi principali, non solo che abbiano oggettivamente gli stessi valori di loudness, ma che siano anche di omogenea qualità rispetto al contesto di fruizione. Ovviamente le due questioni hanno priorità diversa: in prima istanza è necessario che siano allineate le misure di loudness con tutti i necessari controlli di gating o di dialogue intelligence o di quanto altro si sia scelto per operare correttamente la misura di intensità sonora percepita; in secondo luogo è auspicabile che il modello di qualità adottato per la normale produzione di programmi (film, fiction, ecc.) venga esteso a tutti gli oggetti audio compresi gli interstiziali.

4. I SISTEMI E GLI STRUMENTI PER MA MISURA DEL LOUDNESS

Come descritto nel paragrafo 2.1, tutte le misure vengono oramai effettuate sulle rappresentazioni informatiche del segnale. A breve con il passaggio della televisione alle tecnologie DVB-T, ovvero al digitale terrestre, le uniche trasmissioni analogiche saranno quelle radiofoniche, che per altro stanno anch'esse, seppur meno velocemente della televisione,

transitando a tecnologie numeriche. Per tutte le trasmissioni digitali (digitale terrestre, satellitare, iptv, ecc.) è quindi necessario solamente leggere e decodificare il segnale (in realtà questa è un'operazione tutt'altro che semplice o banale, o priva di errori ma non li considereremo in questo contesto) e successivamente operare sul segnale decodificato tutti quegli algoritmi necessari al calcolo del loudness. Possiamo considerare due grandi classi di strumenti, indipendentemente dalle misure e dalle rappresentazioni operanti: una prima classe è quella degli strumenti atti ad un controllo in tempo reale sulle trasmissioni ovvero strumenti che selezionato un determinato canale televisivo o radiofonico, o più semplicemente il suo segnale audio, sono in grado di effettuare una serie di misure su questo segnale in tempo reale; la seconda classe è quella dei sistemi cosiddetti 'file based', in questo caso il segnale deve essere fornito come archivio informatico al sistema che lo analizza e crea un report di misure. Data la potenza di calcolo degli attuali computer le misure di loudness possono definirsi ragionevolmente semplici o comunque tali da essere realizzate con velocità maggiore del tempo reale (ovvero le misure vengono effettate in un tempo minore della durata del segnale). I sistemi file based riescono quindi ad analizzare una maggiore quantità di dati a parità di tempo e sono adatti a effettuare misure su ampi database di segnali o per il controllo del materiale audiovisivo in 'ingest'. I sistemi che lavora invece in tempo reale, che costituiscono la maggioranza dei sistemi in commercio, vengono invece utilizzate negli studi di registrazione, di missaggio e di messa in onda dei programmi e sono un ausilio indispensabile in tutte le regie audio. Spesso, come il caso del sistema della Dolby, questo viene fornito sia nella sua versione di strumento che funziona in tempo reale, sia nella sua versione software.

Figura 6: Il sistema di misura del loudness della Dolby LM100, ed il suo equivalente software

Sebbene esistano ormai sul mercato diversi sistemi che operano misure di loudness (si veda ad esempio oltre a Dolby, TC electronic, RTW, Orban tra i principali) questi non sono spesso così duttili da poter effettuare sperimentazioni comparative. Si è voluto quindi sviluppare un semplice software che realizzasse la misura di loudness su segnali monofonici o stereofonici e implementasse anche la funzione di gating.

4.1 Il software '1770'

Si è quindi sviluppato un programma di consolle in ANCI C che realizzasse tra l'altro la misura di loudness conformemente alla raccomandazione ITU. Il programma opera solamente su segnali audio mono e stereo a 48kHz e rappresentati nel formato Microsoft wav. Il programma, denominato semplicemente '1770' interpreta un file comandi in formato XML, come quello di figura 7.

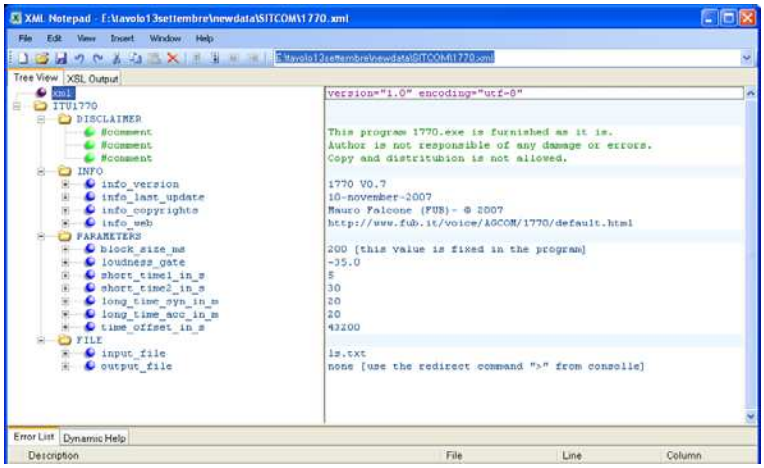


Figura 7: un esempio di file di comando XML per il software '1770'

Il software effettua misure su una lista di file audio, in questo modo è possibile operare misure anche su grandi quantità di segnale senza problemi. Vengono riportate quattro tipologie di misura: il valore RMS, il valore RMS filtrato con il gating, il loudness secondo la norma ITU-R, e lo stesso filtrato con il gating. Queste quattro misure vengono effettuate su tre diverse finestre di misura denominate ST1, ST2, LT1. Queste sono tutte programmabili dal file di controllo e tipicamente si riferiscono a brevi intervalli di tempo (short time) per ST1 e ST2 e a un intervallo di tempo grande, tipicamente da qualche minuto all'ora per LT1 (long time). Infine una quarta finestra di analisi denominata LTA definisce il periodo su cui viene operata una misura integrale che parte sempre dall'istante zero e si aggiorna a multipli di LTA. Infine le quattro grandezze di intensità sonora sopra elencate vengono calcole per ciascun singolo file audio compreso nella lista e per tutta la lista di file audio. Tutti questi risultati vengono forniti in un file Excel la cui prima colonna contiene un identificativo (ad esempio ST1) in modo tale che con un semplice filtro è possibile selezionare le informazioni di interesse. Nel file Excel vengono inoltre riportate tutta una serie di informazioni ancillari e statistiche che per brevità non descriviamo.

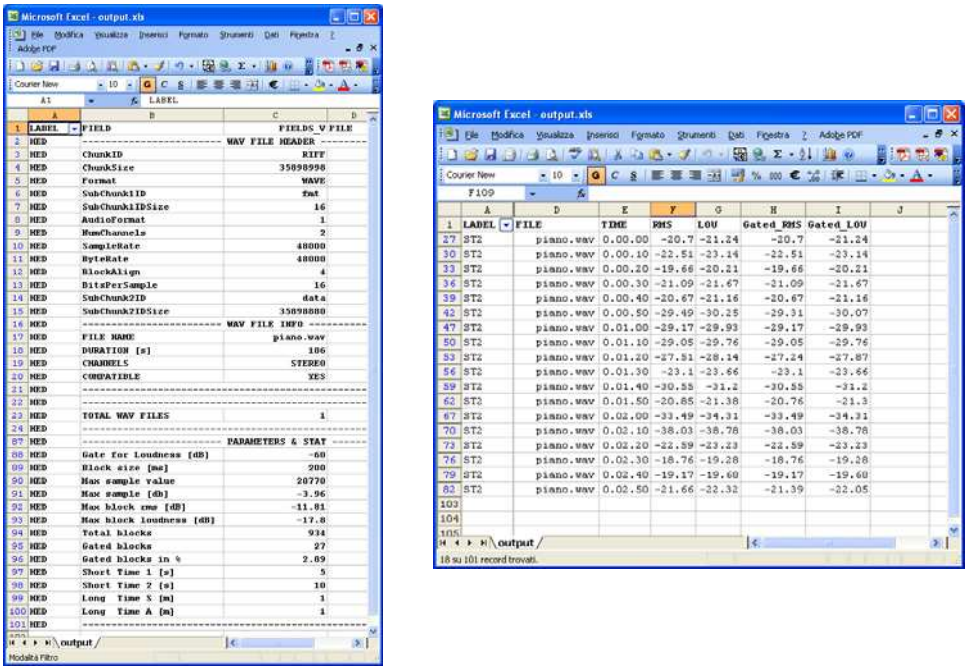


Figura 8: Due schermate relative ai risultati del programma ‘1770’

Da queste è poi semplice realizzare rappresentazioni grafiche di diverso tipo. Tra i vantaggi di un software come questo c’è sicuramente il fatto di poter utilizzare delle liste di file e quindi di poter operare le misure su oggetti audio virtuali di durata illimitata, e dall’altro la velocità di elaborazione molto superiore al tempo reale che permette, su un normale personal computer, di elaborare 12 ore di segnale in circa 10 minuti. La correttezza del software è stata verificata confrontando i risultati ottenuti su alcuni segnali di riferimento misurati con questo e alcuni sistemi professionali in commercio. Tutte le misure riportate in questo lavoro sono state effettuate con il software ‘1770’ e con il sistema Dolby LM100.

5. LA RACCOLTA DEL MATERIALE, IL DATABASE E LE MISURE

Una delle caratteristiche di questo lavoro è l’ampia mole di segnale analizzato. Si è registrato direttamente il segnale audio demodulato in uscita all’interfaccia ottica S/P-DIF (Sony/Philips Digital Interface Format) disponibile in entrambi i ricevitori DVB-T utilizzati: un sistema professionale SENCORE MRD 3187B, ed uno consumer HUMAX DTT5000. Da un’analisi comparativa dei segnali audio acquisiti dai due sistemi, non si è riscontrata alcuna differenza significativa. Quindi, ai fini del nostro lavoro, i segnali audio collezionati con i due sistemi possono considerarsi come provenienti da un’unica sorgente. Il segnale ottico è stato acquisito da un sistema di conversione EDIROL UA5 e memorizzato attraverso una connessione USB direttamente su personal computer.

Uno schema dei sistemi e dell’elaborazione effettuata è riportato in figura 9.

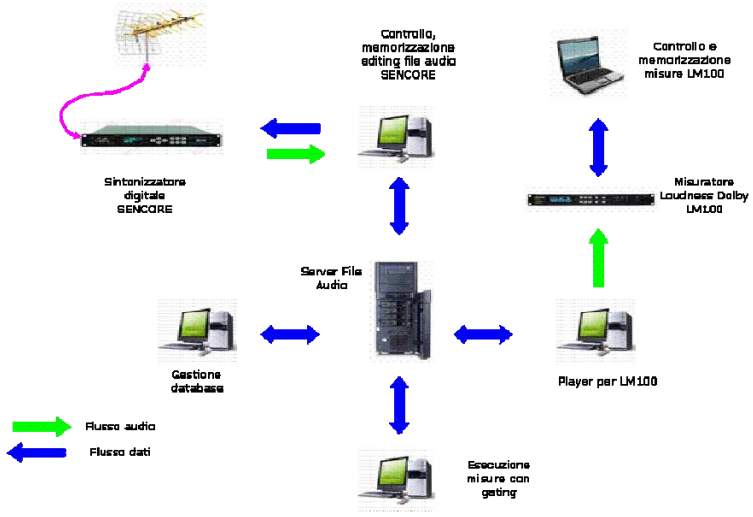


Figura 9: schema di acquisizione ed elaborazione del segnale

Si è registrato il segnale dalle ore 12:00 alle ore 24:00 per tutti i giorni della settimana da lunedì a domenica (in ordine non sequenziale) per 14 diverse emittenti di cui sette di tipo generalista e sette di tipo monotematico. Le registrazioni sono state effettuate in un periodo di circa due mesi, da fine novembre 2008 a inizio gennaio 2009. Avremo quindi un totale di 12*7*14 = 1176 ore di segnale audio analizzato.

Le emittenti registrate sono: RAI1, RAI2, RAI3, RETE4, CANALE5, ITALIA1, LA7 per le generaliste e RAI4, IRIS, RAI-GULP, BOING, MTV, RAI-NEW24, RAI-SPORTPIU perle tematiche.

Per tutte le emittenti generaliste e per i due canali tematici RAI4 e IRIS si è provveduto, sulla base dei palinsesti pubblicati dai rispettivi siti web, dapprima a delimitare manualmente l'inizio di ciascun programma e successivamente a salvare su file audio i singoli programmi di tutte le dodici ore giornaliere. Tali operazioni sono state effettuate con l'ausilio del software Sound Forge della SONY. Per i rimanenti canali, tutti monotematici e con palinsesti giornalieri omogenei, non si è ritenuto necessario operare segmentazione manuale ma solamente un'analisi sincrona ogni 30 minuti.

Un sottoinsieme del materiale proveniente dai canali generalisti è stato segmentato ad un livello superiore separando il programma vero è proprio da tutti i rimanenti contenuti, ovvero da jingle, promo, sigle, spot, spot block, annunci, promozioni, ecc.

Tutto il materiale è stato etichettato seguendo una nomenclatura che permetteva l'immediata identificazione dell'emittente, della data e del nome del programma. Allo stesso tempo, su un'altra postazione, venivano inseriti in un data base tutti i dettagli del tipo di segnale (emittente, sistema di acquisizione, orario indicativo della registrazione, data della registrazione, tipo di programma, titolo del programma) corrispondente al segmento. Questo per permetterne una facile identificazione e rintracciabilità, necessaria per le analisi effettuate sul segnale rispetto alle categorie precedentemente definite (film, fiction, talk-show, TG, ecc.).

L'attuale versione del software '1770' opera la funzione di gating misurando il loudness su una finestra di segnale di dimensione non programmabile da file di comando e pari a

200ms, il livello di gating è invece programmabile e può assumere un valore arbitrario. Tutti i segnali sono stati analizzati con un gating pari a -72LKFS ovvero imponendo un gating che esclude dalla misura del LKFS solo quei segmenti di 200ms con valore inferiore a -72LKFS, in pratica il solo silenzio. Possiamo assimilare questa misura al valore di loudness 'ungated', in quanto si escludono solo i segmenti di silenzio assoluto. Una volta noto il valore LKFS ungated di un determinato segnale possiamo determinare una soglia di gating relativa rispetto a questa, ad esempio imponendo un gating relativo di -8dB andremo a scartare dalla misura tutti quei segmenti di 200ms il cui loudness è inferiore al LKFS ungated meno il valore di gating, ovvero meno otto dB. Oltre alla condizione ungated, si sono operate le misure per livelli di gating pari a -8, -6, -4, -2, e zero.

6. ANALISI DEI RISULTATI

6.1 La variazione 'Channel to Channel'

La prima indagine svolta vuole investigare la differenza tra il livello ordinario di emissione dei vari canali. In questa analisi viene analizzato, in modalità 'blind' ovvero senza considerare nessun tipo di classificazione dei suoi contenuti, tutto il materiale raccolto. Sebbene la stima più accurata del livello ordinario di ciascuna delle quattordici emittenti potrebbe essere computato sul segnale dell'intera settimana, si preferisce riportare i valori giornalieri, dalle ore 12 alle 24, ricordando che i giorni della settimana non sono stati registrati consecutivamente ma randomicamente nel periodo di due mesi come precedentemente descritto. In questo modo è possibile anche quantificare la stabilità di emissione sul lungo periodo di ciascuna emittente.

In figura 10 sono riportati, in un unico grafico, tutti i risultati di questa elaborazione. Ciascun singolo punto nel grafico corrisponde ad una misura di loudness operata su un periodo di dodici ore. Per ciascuna emittente i sette punti collegati tra loro attraverso una linea corrispondono quindi ai valori di loudness dei sette giorni della settimana a partire da lunedì e finendo con domenica, anche se, lo ricordiamo ancora, i giorni non sono temporalmente consecutivi. Le curve in nero (situate sempre nella posizione più in basso del grafico) corrispondono al valore di loudness LKFS ungated. In realtà anche quando si elabora una misura ungated è necessario prevedere l'esclusione dei segmenti di zero assoluto che comporterebbero una divergenza nel calcolo e pertanto si opera un gating molto blando che rimuove solamente i segmenti di zero assoluto. Nel nostro caso quindi il livello per la rimozione dei livelli di zero assoluto è stato arbitrariamente scelto a -72LKFS.

Successivamente, noto il valore di loudness ungated, è stato possibile effettuare le misure di ludness con gating a diversi livelli tutti ovviamente riferiti al valore ungated in nero nel grafico. Ad esempio nel caso si voglia applicare un gating pari a 8dB per il primo giorno della prima emittente, nel grafico il primo punto in nero di RAI1 dal valore di -21.5 LKFS, allora il livello di gating sarà -21.5-8.0= -29.5 e pertanto tutti le finestre di 200ms aventi loudness inferiore a -29.5 non verranno prese in considerazione per il calcolo del loudness con gating -8. La medesima procedura per tutti gli altri livelli di gating sperimentati che ricordiamo essere da -8 a 0 a passo di due, per un totale di cinque diversi livelli di gating. Ovviamente essendo il gating un'operazione che per definizione tende ad aumentare la misura di loudness all'aumentare del gating stesso, avremo che le linee che tracciano per ciascuna emittente l'andamento settimanale del loudness sono tutte ordinate in relazione al gating applicato, ai valori più bassi troviamo sempre in nero le curve per il loudness

ungated, più in alto sempre le curve rosse con il gating più alto, ovvero per un gating pari a zero.

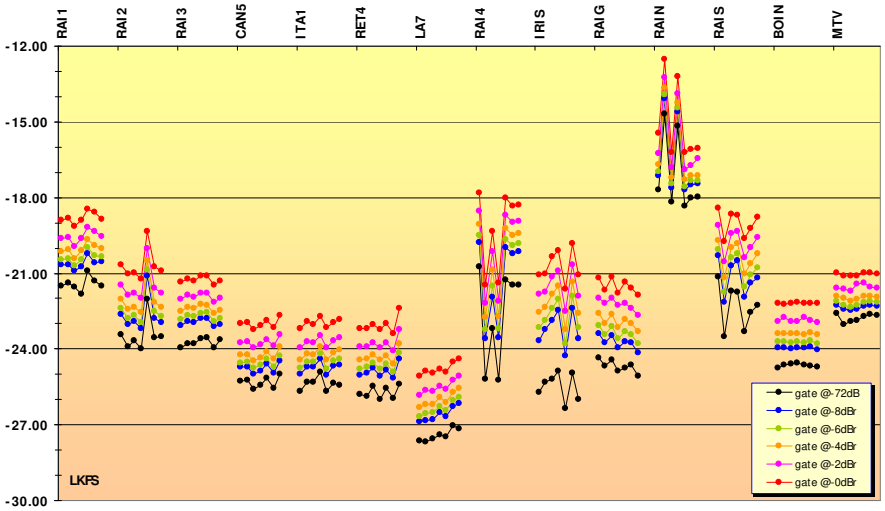


Figura 10: Grafico riassuntivo della variazione 'channel to channel'

Da un'analisi dei dati di figura 10 il fatto più evidente è l'ampia variazione di loudness tra diverse emittenti. Il caso estremo che mette a confronto il loudness de LA7 con quello di RAINEWS24 comporta un divario di loudness, ricordiamo sempre sul valore giornaliero, minimo di circa 9dB e uno massimo di circa 12dB. In altre parole passando dall'ascolto di uno dei due canali all'alto dobbiamo aspettarci (mediamente!) un salto di oltre 10dB.

Un'ulteriore considerazione può essere fatta considerando la stabilità giornaliera del loudness. Si nota infatti come per canali come BOING e MTV il valore di loudness delle varie curve sia praticamente costante durante i sette giorni scelti nello spazio di circa due mesi. Per buona parte delle rimanenti emittenti questa variazione è compresa in un intervallo di un decibel, mentre per pochi casi, RAI4 e RAINEWS24, questa è superiore ai 2/3 decibel. Analizzando questi ultimi casi si vede come questa variazione non sia omogenea ma dovuta a picchi saltuari, come se vi fossero due livelli ordinari di emissione. Un'analisi specifica di ascolto di queste anomalia ha effettivamente mostrato come questi picchi siano dovuti ad un effettiva trasmissione 'traslata' di livello rispetto alle precedenti. Non entriamo ovviamene nel merito delle possibili cause di questo problema, quello che ci preme qui sottolineare che tali variazioni sono legate ad un problema di messa in onda e non di diverso contenuto o allineamento dei programmi.

Infine dobbiamo considerare la variazione introdotta dall'operazione di gating. Come si vede questa è fortemente non lineare, infatti il contributo che il gating comporta comparando il livello ungated (curve in nero nel grafico) con quelle con gating a -8 (curve in azzuro) è di gran lunga minore alla variazione tra le curve con gating a -8 e quelle a gating zero (curve in roso). In generale si nota un raggruppamento dei livelli con gating -8.-6 e -4, mentre gli ultimi due livelli, a -2 e zero, tendono a distanziarsi maggiormente. Per la

maggior parte dei canali l'effetto del gating corrisponde ad un innalzamento della misura di circa 3dB. Tuttavia è chiaro come questo valore dipenda dalla dinamica dei programmi trasmessi. Per canali come MTV e BOING con contenuti, musica pop/rock e cartoon, a bassissima dinamica il gating ha un effetto minore (inferiore ai 3dB). Per canali invece con alta dinamica come IRIS o RAI3, film e fiction, l'effetto risulta ampiamente maggiore. L'effetto del gating infatti è funzione del così detto 'loudness range', ovvero di quel valore che descrive da un punto di vista di vista statistico in quale spazio intorno al suo valore LKFS ungated si muove mediamente il loudness di breve periodo. Audio molto compressi, come quello di musica rock (caso limite è la musica hard rock/metal), avranno un loudness range molto limitato, e l'operazione di gating non produrrà effetti significativi. Infatti in qualsiasi momento ci sintonizzassimo sul una simile musica questo potrebbe indicarsi come rappresentativo dell'intensità sonora. Al contrario se stiamo ascoltando della musica classica o un film ben missato con scene di diverso tipo è molto probabile che la loudness range del nostro audio sia molto ampia in quanto se ci sintonizzassimo casualmente su uno di presti programmi potremmo capitare in un pianissimo, o fortissimo, nel caso della musica classica oppure in una scena molto tranquilla piuttosto che in una scena di azione nel caso di un film. Insomma per avere un'idea più realistica dell'impatto sonoro del programma dovremmo andare a cercare solo quelle scene o brani di musica saliente con volume sostenuto. L'operazione di gating effettua proprio questa selezione fornendo una stima più realistica, o se vogliamo più percettivamente corretta, dell'intensità sonora.

6.2 La variazione 'Program to Program'

Se le variazioni C2C devono considerarsi causate prevalentemente da un disallineamento della catena di messa in onda del broadcaster e solo in minima parte dalla diversità dei contenuti o dal livello del singolo programma, al contrario la variazione 'program to program' è proprio un indice della qualità dei contenuti di un'emittente. Ricordiamo che per effettuare questa misura è stato necessario segmentare il palinsesto giornaliero in singoli programmi. Per alcune emittenti (RAI GULP, RAI NEWS24, RAI SPORT PLUS, BOING, MTV) questa segmentazione del palinsesto risultava molto difficile e pertanto si è preferito effettuare un'analisi sincrona segmentando ogni trenta minuti. Dobbiamo pertanto caratterizzare la variabile corrispondente al loudness del programma per tutte le emittenti selezionate e su tutto il periodo di registrazione. Per semplicità si è scelto di rappresentare solamente il valore di loudness con gating a -8dB. In questo caso, diversamente dal precedente e dalla maggior parte dei casi, non è facile realizzare una significativa rappresentazione mostrando i risultati delle misure, è piuttosto necessario rappresentare alcune grandezze statistiche della serie di misure ottenute. Con un rappresentazione come quella di figura 11 è possibile racchiudere in un unico grafico un'esauriente descrizione dei risultati. Una volta calcolato il loudness di tutti i programmi della settimana se ne calcola la sua distribuzione e da questa è possibile calcolare il percentile, ovvero quel valore di loudness tale che una determinata percentuale di programmi abbia volume superiore al predeterminato valore. Nel grafico seguente per ciascuna emittente si è rappresentato: il valore mediano di loudness dei programmi che corrisponde anche al valore percentile del 50% ed è rappresentato nel grafico dal segmento orizzontale all'interno del box rettangolare; il box rettangolare rappresenta invece il percentile al 25% (lato inferiore del box) e al 75% (lato superiore); infine gli estremi della linea verticale che taglia il bos corrisponde ai valori del programma di massimo e di minimo loudness. Consideriamo, ad esempio, il caso di RAI4. Nell'arco delle registrazioni effettuate il loudness del programma a minor volume è di -

25.8LKFS (estremo in basso della linea verticale), mentre il programma a volume più sostenuto ha raggiunto il valore di -17.6LKFS. La mediana del loudness di tutti i programmi, ovvero il valore per cui metà dei programmi hanno valore superiore e l'altra metà inferiore, è di -21.1LKFS (segmento orizzontale all'interno del box). Infine avremo che il 50% dei programmi (ovvero il percentile che va dal 25% al 75%) ha un loudness compreso tra -23.0 e -19.9 LKFS che sono appunto la dimensione del box.

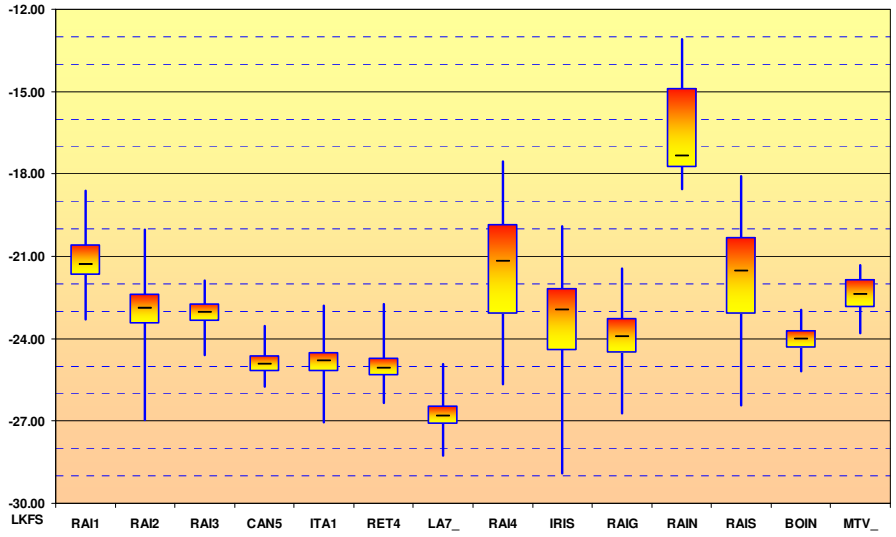


Figura 11: rappresentazione percentile e min-max della variazione 'program to program'

Se consideriamo che nel caso ideale in cui ogni programma fosse allineato ad un ben definito valore di loudness, cosa per altro tecnicamente possibile senza grandi difficoltà, tutto questo grafico dovrebbe condensarsi nei soli tratti, per altro tutti allineati, rappresentanti la mediana in quanto tutte le altre grandezze statistiche verrebbero a coincidere con questa. Questa situazione, come abbiamo detto sicuramente realizzabile e che probabilmente sarà la base di una raccomandazione da parte degli organismi regolamentari internazioni, rappresenta il caso di qualità ideale per quanto riguarda l'allineamento del loudness.

Dai risultati sperimentali invece si evince come la situazione nel nostro paese sia particolarmente scadente e come tra i programmi di loudness minimo e quello massimo vi sia una differenza di ben 16dB tra diverse emittenti, o di 9dB all'interno di una stessa emittente.

Tuttavia i dati risultano coerenti in quanto mostrano, ad esempio, come emittenti della medesimo broadcaster e con caratteristiche di programmazione simile hanno figure simili come nel caso dei tre canali principali Mediaset. Analogamente si evince come emittenti come IRIS e RAI4 hanno una distribuzione del loudness molto meno concentrata dei rispettivi canali generalisti dei rispettivi broadcaster, probabilmente perché nei film e nelle fiction trasmesse non viene effettuato un preventivo riallineamento del materiale in ingest

che verrà trasmesso. Ma questa è ovviamente solo una nostra ipotesi che cerca di spiegare, e non certo di giustificare, un panorama piuttosto desolante per quanto riguarda lo stato della qualità audio e del loudness nel nostro paese.

6.3 Ulteriori considerazioni sulle misure di loudness

Il terzo e certamente più discusso caso di disallineamento del loudness è quello relativo alle inserzioni pubblicitarie ovvero al 'program to advertising'. Non discuteremo esplicitamente qui di questo problema già trattato in un altro lavoro (Falcone, 2006) sia perché particolarmente complesso, sia perché negli ultimi anni vi è stato un fervente proliferare di normative a riguardo. Più interessante ci sembra invece confrontare l'efficacia della tecnica di gating con altre metodologie come quella del 'dialogue intelligence', e specializzare tale confronto in diversi tipi di contesti e per i diversi contenuti della programmazione. Anche in questo caso, per quanto riguarda il gating si è scelto di operare a un livello di -8dB rispetto al valore LKFS ungated. Per quanto riguarda invece le misure sul dialogo si è utilizzato lo strumento LM100 della Dolby. Poiché detto strumento lavora in tempo reale la quantità di segnale utilizzato in queste analisi non può essere esaustiva di tutto il materiale raccolto come nei casi precedenti.

Come prima cosa compariamo queste metodologie in modalità blind ovvero su tutto il segnale trasmesso nell'arco delle 12 ore. Abbiamo preso a campione una registrazione di 12 ore per una serie di sei emittenti con diverse caratteristiche di contenuti, e su queste abbiamo effettuato le tre misure: con lo strumento LM100 (barra verde in figura 12), di loudness LKFS ungated (punti blu), e di LKFS con gating a -8dB (punti rossi).

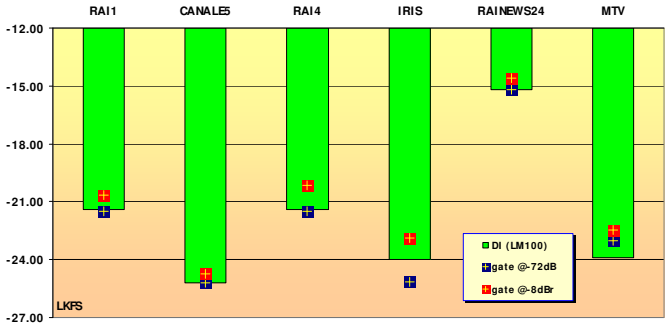


Figura 12: Comparazione del loudness nel caso di trasmissione giornaliera

Come si vede nei canali generalisti e in quelli prevalentemente con segnale parlato le tre misure sono praticamente coincidenti o comunque entro 1dB. Per le emittenti a contenuti prettamente legati film o fiction si vede come il livello del parlato non sia equivalente a quello ottenuto con il gating ma sia leggermente più basso di circa 1dB. Anche per l'emittente musicale MTV il livello del parlato risulta di poco più di 1dB più basso del loudness misurato con il gating. In conclusione potremmo dire che in generale le due metodologie sono compatibili solo se il contenuto della programmazione è prevalentemente parlato (cosa ovviamente facilmente presumibile), mentre per altre tipologie sembrerebbe, almeno nei limitati casi presi in considerazione, che la tecnica utilizzata dal LM100 tende a dare valori più bassi rispetto a quella del gating. Più bassi di circa un decibel con un gating molto blando di -8dB, ma che possono, anche per quanto precedentemente mostrato,

raggiungere valori ben maggiori con livelli di gating più stringenti o nel caso di contenuti particolari.

I film costituiscono certamente un buon banco di prova. Abbiamo pertanto selezionato quattro film da quattro diverse emittenti e abbiamo manualmente diviso il contenuto di quanto trasmesso in tre parti: A, ovvero il film vero e proprio senza alcuna contaminazione; R, ovvero tutti quei segmenti relativi a annunci, spot, promozioni, jingle e quanto altro vengono trasmessi durante il film; S, ovvero una selezione operata manualmente, e arbitrariamente, da un operatore esperto che seleziona solo quei passaggi del film in cui l'audio corrisponde alla sola voce senza musica o rumori di fondo. Le misure in questo caso mostrano come per tutti è tratto i film delle quattro emittenti nel caso dei segnali interstiziali (pubblicità, promo, ecc.) le tre misure vengano a coincidere confermando che questi segnali sono di scarsa dinamica o comunque con un loudness range molto limitato, come mostrato in figura 13. Completamente diverso è il caso per l'audio dei film dove il valore di LKFS con gating è significativamente minore di quello del LM100, e comunque sempre ben distanziato dal LKFS ungated. Nel caso si analizzi solo il segnale di parlato puro, individuato da un operatore esperto, i valori di LKFS ungated e del LM100 sono pienamente compatibili, come avremmo dovuto aspettarci, ed anzi una selezione manuale, sicuramente più accurata di quella operata dal LM100, mostra valori di loudness anche leggermente minori rispetto a questi che già, come nel caso precedente, risultano sempre più bassi di quelli del LKFS con gating a -8dB.

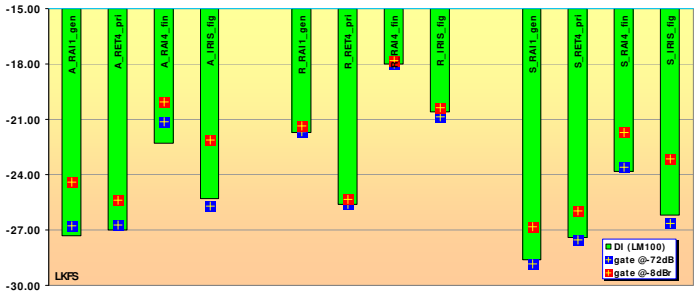


Figura 13: Comparazione del loudness nel caso di film

Si evidenzia quindi, ancora una volta, come i contenuti interstiziali, assumano caratteristiche completamente diverse rispetto ai loro contenitori quando questi sono programmi di buona qualità audio come nel caso dei film da noi selezionati. È quindi ragionevole aspettarsi, viste anche le diverse caratteristiche o se vogliamo la diversa qualità dei segnali audio, che questi possano a parità di loudness avere un impatto percettivo diverso.

La medesima analisi è stata operata dopo aver selezionato quattro fiction di qualità, di recente produzione. Anche in questo caso si è considerato il solo programma (A), l'insieme dei segnali interstiziali (R), e il segnale selezionato da un operatore esperto che contiene solo parlato privo di rumori o musica (S). Anche le misure effettuate sono le medesime dell'esempio precedente. I risultati sono riportati in figura 14. Essendo film e fiction di recente produzione due tipologie di programmi quasi identici come tecniche di produzione e come contenuti, non dobbiamo sorprenderci della quasi identità dei risultati ottenuti nei

due casi. Quanto detto per il caso precedente vale infatti per i risultato ottenuti per le fiction. L'unica minima differenza e che ora la differenza tra LKFS ungated e gated è, per il segnale di tipo A, ovvero per la fiction vera è propria, ed in parte per il solo parlato S, mediamente e leggermente minore rispetto a quello dei film. Questo ovviamente è del tutto comprensibile e giustificabile se consideriamo il fatto che la qualità delle fiction è più vicina al mercato televisivo che a quello cinematografico.

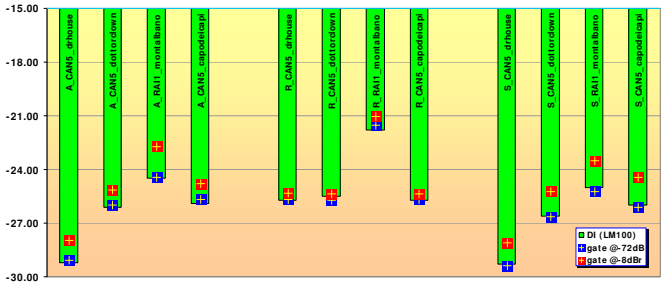


Figura 14: Comparazione del loudness nel caso di fiction

L'ultimo esempio invece si riferisce a programmi tipo talk show per lo più trasmessi in diretta televisiva. I risultati sono riportati in figura 15. In questo caso ci aspetteremmo una ancora più stretta concordanza tra le diverse misure, e di fatto questo è proprio quello che si ottiene. Per tutti e quattro i talk show selezionati, e per tutti e tre i tipi di segnale, A, R e S le diverse misure risultano compatibili entro circa 1dB. Questo significa che in questo caso la misura del loudness operata con il 'dialogue intelligence', con il semplice LKFS o con il gating producono all'incirca il medesimo risultato.

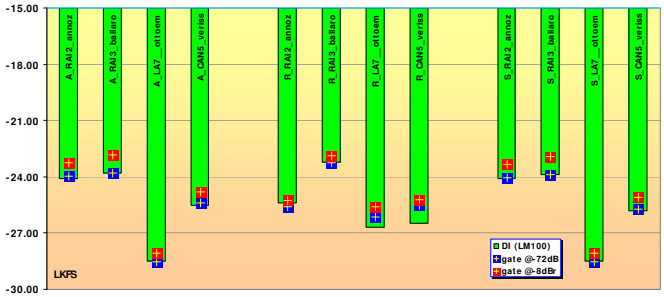


Figura 15: Comparazione del loudness nel caso di talk show

Ma questo è proprio quello che ci aspettavamo, infatti in questo caso il programma è caratterizzato proprio dal parlato (non a caso i valori per A e S sono praticamente identici) e il gating non ha significative zone di segnale da rimuovere in quanto il parlato è mediamente tutto allineato a valori significativi di loudness.

Ripercorriamo ora l'analisi delle figure 13, 14 e 15, osservando i risultati da un altro punto di vista. Confrontiamo per ciascuna analisi i valori di loudness di un programma relativamente al suo contenuto principale (segnale A) con i valori di loudness degli interstiziali ovvero essenzialmente della pubblicità (segnale R). Nel caso di film si vede come il segnale R sia nettamente superiore a quello A, in alcuni casi anche di più di 3dB. Nel caso

di fiction vediamo invece che per due casi (del medesimo broadcaster) il segnale R risulta nettamente superiore a A, mentre negli altri due (di un medesimo ma diverso dal precedente broadcaster) si equivalgono. Infine nel caso di talk show ci troviamo in una condizione totalmente diversa dove sono verificate tutte le condizioni anche se il divario tra A e R non è mai particolarmente significativo come nel caso dei film. Questo significa che quanto dettato per legge e riassunto nel paragrafo tre, è ancora ampiamente disatteso specialmente se consideriamo un contesto a breve termine ovvero il confronto del livello della pubblicità con il livello del programma che lo contiene, che è, lo ricordiamo, l'unico tipo di confronto significativo e che per altro anticipa quanto si sta elaborando nei diversi tavoli normativi internazionali.

7. LOUDNESS WAR E SUO IMPATTO SULLA QUALITÀ DEL SEGNALE AUDIO E DEL SEGNALE VOCALE

"L'espressione 'loudness war' (o 'loudness race'), in italiano traducibile in 'guerra del volume', si riferisce alla tendenza dell'industria musicale a registrare, produrre e diffondere musica, anno dopo anno, con livelli di volume progressivamente più alti, per creare un suono che superi in volume i concorrenti e le registrazioni dell'anno precedente" (Wikipedia, 2008). Come tutte le guerre, anche la loudness war ha prodotto e continua a produrre vittime anche al di fuori del suo contesto originale del mercato musicale (Sreedhar, 2007). Infatti, la medesima tendenza la ritroviamo in tutti gli scenari audio e quindi anche in quelli radiotelevisivi. La ricerca di tecniche e strategie per ottenere un volume più forte come fattore vincente, sembra essere dilagata a scapito della qualità e della fedeltà dei segnali. Come abbiamo detto un impatto sonoro più forte è possibile averlo anche a parità di misure oggettive. Per questo con raccomandazioni come la ITU-R BS.1770, e con l'introduzione di tecniche come quelle del gating si cerca di ridurre al minimo la possibilità di avere, a parità di misura oggettiva, segnali di diverso impatto percettivo. Tuttavia le moderne tecniche di elaborazione del segnale permettono manipolazioni, ed in particolare compressioni di segnale che riescono, in parte, a eludere le semplici misure oggettive di loudness. Ma questo è solo un danno marginale. Il problema, a nostro avviso più grave, è che tale tipo di attacco alla qualità del segnale audio si è velocemente diffuso in tutti gli ambiti producendo di fatto un nuovo standard di riferimento a cui il pubblico si è assuefatto, e che ora non solo subisce passivamente ma addirittura può preferire all'originale proprio perché indotto ad una cattiva cultura dell'esperienza audio. Così un segnale sovracompresso e del tutto innaturale non solo più essere preferito, ma può erroneamente essere scambiato per quello più aderente alla realtà e pertanto fornire un modello di riferimento sbagliato, o comunque certamente artificiale, nel nostro apprendimento attraverso l'esperienza. Per quanto riguarda il parlato questo non fa eccezioni. Se la televisione in passato è stato uno strumento per unificare la lingua, oggi potrebbe divenire lo strumento per snaturalizzare, o se non altro per modificare secondo altri canoni, lo stile prosodico del parlato e come prima cosa il controllo della intensità ovvero del loudness. Se da un lato può essere facile verificare come la televisione e la radio hanno 'insegnato' a parlare una lingua condivisa a tutti gli italiani quando la fruizione giornaliera di questo mezzo era molto limitata rispetto a oggi, l'impatto che questa ha oggi nell'influenzare le modalità di produzione acustica nel parlato è un campo poco o per nulla studiato. Certo è che il modello che oggi viene proposto corre il rischio di diseducare piuttosto che di

educare. Nell'ambito musicale questo è certamente vero, ma visti i presupposti non è difficile supporre che questo accada anche per il parlato.

Non pretendiamo qui di effettuare uno studio analitico e di caratterizzare il problema, ma solo di evidenziarne l'esistenza. A semplice titolo di esempio vogliamo riportare un caso estremo, relativo al segnale di una fiction di recentissima produzione e il cui audio è stato curato da provati professionisti. In figura 16 è rappresentata la forma d'onda del segnale relativo ad un normale dialogo, e ad un fortissimo urlo entrambi pronunciati dalla stessa persona e in due scene aventi in primo piano l'attore. L'innaturalezza della cosa è palese già alla semplice analisi della forma d'onda. Il loudness del segnale urlato è di circa 2/3 dB superiore a quello del segnale parlato, nonostante l'ampiezza della forma d'onda sia inferiore a quella del parlato ma di questo effetto dovuto alla compressione ne abbiamo già discusso. In natura una voce urlata dovrebbe essere di almeno 15 dB superiore rispetto ad un normale parlato. Certamente possono esservi delle necessità o scelte sceniche per avere delle realizzazioni come quelle della figura, ma all'ascolto la voce urlata comunque viene effettivamente percepita come tale e di intensità corretta nel contesto, sia per le sue caratteristiche, sia per effetto delle manipolazioni e della compressione del segnale. Di fatto costituisce un verosimile riferimento, e pertanto un esempio, di un evento vocale in cui l'intensità gioca certamente un ruolo fondamentale.

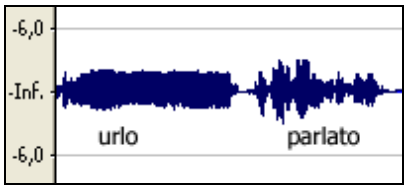


Figura 16: Due esempi di voce a confronto in una fiction televisiva

Infine una cautela ulteriore deve essere espressa nell'utilizzo dei molti data base vocali raccolti dalla televisione. Il fatto che la qualità audio sia migliore di quella telefonica non deve indurci a poter considerare questo segnale come un riferimento per lo studio del parlato naturale che rispecchia una situazione di fedeltà. Il segnale televisivo è pur sempre un segnale 'artificiale' in quando risultato di una tecnica di missaggio che ha il compito di dare voce ad un determinato oggetto scenico, deve cioè rispecchiare le caratteriste e i requisiti di un obiettivo artistico piuttosto che tecnico e difficilmente, anzi praticamente mai, sarà copia fedele della realtà. Ma come abbiamo detto questo è solo uno degli aspetti della manipolazione, un secondo e più subdolo fattore, in quanto non derivante da scelte di operatori e/o tecnici audio, è quello dovuto ai sistemi di compressione e di allineamento automatico utilizzati comunemente nella messa in onda dei programmi. Il risultato finale è che ciò che arriva nelle nostre case non solo non è copia fedele della realtà, ma non è neppure esattamente quello che i tecnici audio con fatica e professionalità avevano realizzato per dare voce al programma. Anche al fine di eliminare questa pessima consuetudine, diversi gruppi di lavoro stanno cercando di modificare le norme e la regolamentazione in modo di avere una catena di elaborazione audio, dal produttore all'utente finale, il più semplice e chiara possibile.

8. ATTIVITÀ NORMATIVE SUL LOUDNESS

Ormai possiamo dire con certezza che la misura di intensità sonora sta in questi anni vivendo una completa e nuova rifondazione. Dopo le prime indicazioni formulate con la raccomandazione ITU-T BS. '1770', oggi siamo probabilmente al picco di attività in tale ambito sia per quanto riguarda le attività normative, sia per quanto riguarda i prodotti sul mercato, e le aziende più virtuose stanno già adattando i loro protocolli di produzione e controllo della qualità in tal senso. A livello internazionale l'ITU sta già lavorando su una nuova versione della raccomandazione '1770' che includa anche la misura del LFE per i segnali multicanali, e che utilizzi la tecnologia del gating per misurare quello che viene indicato come 'foreground loudness' e che deve considerarsi come il riferimento percettivo di un programma. Un fatto importante è che tale misura del loudness è già proposta per rimodulare tutte le altre raccomandazioni che fanno uso di vecchie metodologie di misura. Pertanto anche la raccomandazione internazionale che definisce il livello audio standard nello scambio di programmi (anche attraverso diverse emittenti) sta per essere riformulato e sarà probabilmente definito come il livello LKFS dell'intero programma con un determinato valore di gating. Cosicché ogni qual volta si deve scambiare un programma, film, fiction o anche un programma in diretta televisiva come una partita o altro, l'emittente distributrice sarà tenuta a fornire il programma in modo che il suo livello LKFS con gating sia un ben determinato valore (probabilmente qualcosa intorno ai -24LKFS). Quando sarà definita tale norma è facile pensare che la messa in onda dei programmi possa avvenire senza alcun strumento intermedio che ne regoli, o meglio ne alteri, il volume in quanto tutti i programmi dovranno essere normalizzati in loudness. Questo risolverebbe in un unico colpo tutti i problemi descritti in questo lavoro, e se applicata a ogni tipo di produzione, ovvero anche agli spot pubblicitari, risolverebbe, sicuramente in modo migliore di come gli organi preposti stanno faticosamente facendo, anche il problema della pubblicità. Insomma con un colpo solo tutti e tre i disallineamenti descritti in questo lavoro potrebbero essere risolti. Tuttavia determinare i parametri ottimali per il gating, definire una misura unica per tutte le tipologie di segnale (voce, musica, ecc.) e per tutte le modalità di distribuzione (mono, stereo, multicanale) non è cosa facile. Per questo motivo lo scorso anno l'EBU (European Broadcaster Union) ha creato uno speciale gruppo di lavoro, denominato P/LOUD, con il compito di risolvere questi problemi, anche a supporto del'ITU, e comunque con il fine di determinare, diffondere e promuovere una raccomandazione di categoria approvata da tutti i broadcaster. La partecipazione a tale gruppo è maggiore di quanto fosse originariamente previsto e i lavori procedono speditamente anche grazie all'altissimo livello di professionalità del gruppo che è costituito dagli esponenti di punta per quanto riguarda il controllo della qualità audio di tutti i principali broadcaster europei, con contributi anche da broadcaster internazionali. Infine vi è il caso nazionale dell'Autorità Garante delle Comunicazioni che ha il compito di controllare che la legge, già citata nel paragrafo tre, sia rispettata e pertanto ha il compito di definire le modalità operative di misura. A tal fine l'Autorità ha emesso una prima delibera nel 2006 definendo una metodologia di misura. Nel 2007 costituisce un tavolo con le parti per rivedere tale metodologia e nel 2009 emette una seconda delibera, e contestualmente apre un secondo tavolo di consultazione. La vigente delibera del 2009 prevede l'utilizzo della misura secondo la raccomandazione ITU BS.1770 con l'introduzione di un gating a -8dB, su finestre di segnale di mezzo secondo. In qualche modo quindi precorre i tempi e si pone come una soluzione fortemente il linea con le più moderne raccomandazioni e tendenze.

Dall'altro, tuttavia, le scelte operate per le modalità di confronto tra programma e pubblicità decontestualizzano fortemente i due termini, e non prevedono un allineamento generale, come invece le nuove raccomandazioni ITU sembrano suggerire. Tale decontestualizzazione, oltre a complicare inutilmente con una serie diversa di stime del livello ordinario della programmazione, non impone di fatto un allineamento tra programma e pubblicità che rimane, a nostro avviso, l'unica e cardinale questione da risolvere sulla base delle misure qui discusse.

9. CONCLUSIONI E FUTURE ATTIVITÀ

In questo lavoro si è ampliato e ripreso il problema già affrontato precedentemente (Falcone, 2006). Si è analizzata una quantità di dati tra le più ampie mai utilizzate a livello internazionale, e sicuramente la più ampia tra quelle riportate in letteratura per il segnale televisivo nel nostro paese. Con un totale di oltre mille ore si segnale analizzato, si sono potuti investigare, in maniera statisticamente significativa, tutti i diversi tipi di disallineamento del loudness. Attraverso le misure più moderne, e mettendo a confronto le metodologie più promettenti per la stima del cosiddetto 'foreground loudness' si è mostrato come la situazione attuale rispecchi uno scenario particolarmente caotico e particolarmente irrispettoso rispetto agli utenti in quanto affatto attento alla qualità del segnale audio.

È nostra ferma convinzione, e i lavori di frontiera in questo ambito lo confermano, che sebbene una corretta definizione del loudness ed un suo utilizzo nella produzione e diffusione del segnale audio e della voce siano un cambiamento epocale nell'ambito delle tecnologie audio, questo sia solo il primo passo verso la soluzione di un problema, quello della qualità audio tout court, che è ancora lontano dall'essere risolto. Misure statistiche di ordine superiore sul livello del segnale, piuttosto che pesature che tengano conto del contenuto e delle sue caratteristiche audio, e infine l'influenza del sistema e della modalità di riproduzione costituiscono tutti fattori che solo in prima approssimazione possono essere uniformati, controllati e ottimizzati con una singola grandezza come si sta facendo con le misure di loudness discusse in questo lavoro. Ciò nonostante si è dimostrato come anche con una prima approssimazione di queste misure ovvero con il semplice loudness, se opportunamente coadiuvato ad esempio con strategie di gating e se correttamente utilizzato nella produzione e distribuzione dei programmi, sia possibile risolvere semplicemente ed efficacemente i molti spiacevoli problemi evidenziati nell'analisi del materiale raccolto. Estendere perciò la ricerca e la sperimentazione verso parametri globali di qualità audio, dato che la molteplicità dei mezzi e della tipologia dei segnali va sempre più diversificandosi, non è solo una necessità tecnica imprescindibile, ma anche una sfida intellettuale che la comunità scientifica dell'audio e della voce deve affrontare. Ed è su questa linea che intendiamo portare avanti i nostri studi, ed in particolare gli aspetti ed il ruolo che il segnale vocale comporta nei diversi contesti e scenari.

RINGRAZIAMENTI

Si vuole ringraziare tutto il personale dell'ISCOM e della FUB che ha supportato le difficili fasi di acquisizione ed elaborazione del segnale audio televisivo. Si vuole inoltre ringraziare tutti i membri che hanno partecipato insieme agli autori ai vari tavoli nazionali e internazionali sul problema del loudness, e che sono stati fonte di informazioni e di esperienze relative a chi lavora dalla parte delle televisioni.

10. BIBLIOGRAFIA

Chiocci, F., Cordoni, G., Ortoleva, P., Sibilla, G., (2002), *La grana dell'audio. La dimensione sonora della televisione*, Comunicazione/VQPT – Verifica qualitativa programmi trasmessi (VQPT 188), Roma: RAI-ERI.

Falcone, M., Barone, A., Bonomi, A., Monaco, G., Ciavatta, D. (2006), Abbassa quello spot per favore, in *Scienze vocali e del linguaggio. Metodologie di valutazione e risorse linguistiche,* Atti del 3° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Trento, 29 novembre-1 dicembre 2006 (V. Giordani, V. Bruseghini & P. Cosi, editors), Torriana: EDK Editore, 321-339.

ITU-R, Recommendation BS.1770 (2006), Algorithms to measure audio programme loudness and true-peak audio level, Geneva, Switzerland.

Scopece, L. (2009), *L'audio per la televisione*, Roma: Gremese Editore.

Sreedhar, S. (2007), The Future of Music, *IEEE Spectrum* online, August 2007, http://www.spectrum.ieee.org/aug07/5429.

Wikipedia, (2008), definizione della voce '*Loudness War*' su Wikipedia, la libera enciclopedia, http://it.wikipedia.org/wiki/Loudness_war.

SONORITY BASED SYLLABLE SEGMENTATION

Bogdan Ludusan, Serena Soldo
Department of Physical Sciences, 'Federico II' University, Naples
ludusan@na.infn.it, soldo@na.infn.it

1. ABSTRACT

This paper proposes a new method for detecting syllable boundaries. It is based on the sonority and it uses the so-called 'Sonority Sequencing Principle' for the boundary detection. As acoustic correlate of the phonological concept of sonority we use the regularities present in the spectrogram of the signal. By finding the maxima of the sonority function we will be finding the syllable nuclei, while the syllable boundaries are to be found at the minima of the sonority function. Due to the fact that it uses only the information contained in the speech signal it could be implemented, with small modifications, for almost any language.

2. INTRODUCTION

Automatic speech segmentation is a topic of great interest in nowadays speech related literature due to its multiple use. One of its most important application areas is Automatic Speech Recognition (ASR), in which speech segmentation techniques are applied for obtaining the units used for recognition. In the recent ASR literature, the syllable is a frequent choice for such a unit because it offers a good representation of the variability present in the speech signal while retaining a also good trainability. This is the reason behind our proposal for an algorithm for automatic syllable segmentation.

Although the syllable is intuitively recognized by most of the people, there is not yet a universally agreed definition of the syllable. For example, from an acoustic point of view it was observed that energy temporal patterns play a fundamental role (Jespersen, 1904), syllable nuclei being usually found in correspondence with energy maxima, while syllable boundaries correlating with energy minima. In contrast, in phonology, the most widely used syllable definition is based on the sonority scale.

The sonority is a concept present in the phonological theory from the nineteenth century. The opinions on whether the sonority has or not a phonetic basis are divided some suggesting that it is correlated in some way with audibility (Sievers, 1881), some that it can be defined in terms of the loudness of a sound, which is related to its acoustic energy relative to other sounds having the same length, stress and pitch (Ladefoged, 1993), while others do not even recognize it as a phonological concept (Harris, 2006). Taking on a different stance, Clements (1990) argues that the absence of a physical basis for characterizing sonority in language-independent terms would make it impossible to explain the nearly identical nature of sonority constraints across languages.

Based on the measure of sonority, several relative rankings of the sonority of sounds were developed, among which, I recall the one presented in (Ladefoged, 1993): low vowels > mid vowels > high vowels > liquids > nasals > obstruents. The Sonority Sequen-

cing Principle (SSP) is used as principle for syllabification stating that the sounds inside a syllable increase in sonority from the onset to the nucleus, with a maximum value corresponding to the nucleus and decrease in sonority from the nucleus to the coda.

The sonority was used previously as feature for segmentation in speech processing, but it was either used to detect only syllable nuclei (Kawai & van Santen, 2002) or to detect syllable boundaries, but combined with other features and in conjunction with statistics from previous segmentations (Mayora-Ibarra & Curatelli, 2002).

In (Kawai & van Santen, 2002) multiple linear regression is used in order to obtain, what the authors call, the instantaneous sonority. As predictor variables for the regression they use bandpass-filtered acoustic energy from the central part of each phone. The authors argue that the five frequency bands chosen can efficiently locate boundaries between different phone classes. They report accuracies of over 60% for syllable nuclei detection and over 80% for speech rate recognition for a corpus of read news.

Mayora-Ibarra & Curatelli (2002) obtain their segmentation by using time-domain signal processing followed by a refinement of the results based on a fuzzy-logic approach. As time domain feature they use the zero-crossing rate in the intervals of sonority decrease, which, they state, it is related to the attenuation of the acoustic intensity of speech that occurs between the transition of adjacent syllables. The second step represents a refinement of these results and it is implemented using statistics from previous segmentation tests together with fuzzy logic rules. The accuracies reported on a corpus of isolated Italian digits are of 87% after the first phase and 95% after the refinement of the results.

Recent work (Galves *et al.*, 2002) has proved the usefulness of the sonority in other areas, like rhythmic class discrimination. In their paper, the authors propose a formulation for the sonority function, defined on the interval [0,1]. The proposed function has values close to 1 for sounds displaying regular patterns, characteristic of sonorant portions of the signal and close to 0 for regions characterized by obstruency.

In Cassandro *et al.* (2002) the authors refine the previously proposed function using an exponential. Subsequently, the sonority is defined as a decreasing function of the values of the relative entropies between neighbouring columns of the spectrogram of the speech signal:

$$S(t)=\exp(-\beta\sum_{1}^{3} h(p_t|p_{t-i})) \qquad (1)$$

where h denotes the relative entropy between two probability measures, p_t is the power spectrum renormalized in order to become a probability measure and β is a free parameter assuming positive real values.

Among the methods used in the literature for syllable boundary detection there are many algorithm using only the information extracted from the speech signal, without any linguistics or phonetic knowledge (Petrillo & Cutugno, 2003; Nagarajan *et al.*, 2003). The first approach (Petrillo & Cutugno, 2003) is based on the energy of the signal and it searches the syllable boundaries at the minima of the energy envelope. In Nagarajan *et al.* (2003), the authors obtain the syllable segmentation based on a minimum phase group

delay approach. To our knowledge, the results presented in the previous paper are the best segmentation results on an English corpus of conversational speech.

3. METHODS

3.1 Algorithm

Based on the previous formulation of the sonority function (1), we propose an algorithm for the detection of syllable boundaries. The algorithm uses exclusively speech processing techniques (both frequency and time domain), having no knowledge about the phonetic content of the signal, in order to obtain the syllable boundaries from the continuous speech signal. Figure 1 presents the block scheme of the algorithm.

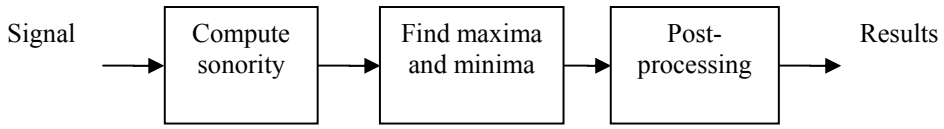


Figure 1: Block scheme of the algorithm

In a first step, for each of the utterances, the sonority function is computed in a similar manner to the one described in Cassandro *et al.* (2002). The major difference consists in the use of a different distance function for the computation of the sonority – the normalized Euclidean distance instead of the relative entropy between the columns of the spectrogram. The following steps are executed in order to obtain the sonority of the signal:

1. computing the spectrogram of the signal for the frequency band below 1000 Hz, using a 25 ms window;
2. normalization of the power spectrum;
3. computing the normalized Euclidean distance of five consecutive columns; the formula of the normalized Euclidean distance between two vectors (here the vectors representing the columns of the spectrogram) is listed in (2); the distance function used has the same properties as the relative entropy – gives low values for vowels, nasals and voiced stops (because of the regularity introduced by voicing) and high values for voiceless stops, fricatives and flaps (Garcia *et al.*, 2002):

$$d(\vec{x},\vec{y})=\sqrt{\sum_{i=1}^{p}\frac{(x_i-y_i)^2}{\sigma_i^2}} \qquad (2)$$

where σ_i is the standard deviation of x_i over the sample set;

4. applying relation (1) for the normalized Euclidean distance we will obtain the value of the sonority function; by multiplying in the exponential function with -1, we will obtain in the sonority function high values for vowels, nasals and voiced stops and low values for all the other sounds.

As an example, the sonority profile of the word ‘cinquecentoventunomiladuecentouno’ along with its syllable segmentation is presented in Figure 2.

The frequency band for which the spectrogram of the signal is computed (0-1000 Hz) was established empirically. The bandwidth was determined through testing of various bandwidths between 400-500 Hz (in order to catch at least the F0 of the voiced segments) and 1500-2000 Hz (to avoid a high computation time for the sonority function).

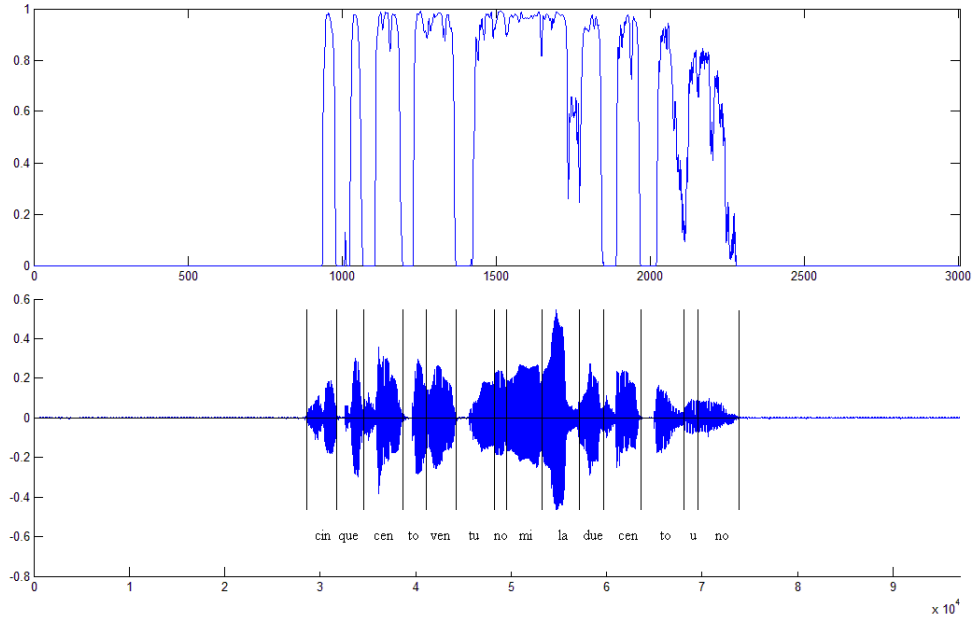


Figure 2: Sonority profile and syllable boundaries for the word 'cinquecentoventunomiladuecentouno'

As the distance metric used is the most important factor in computing the sonority function, we needed a robust function for it. By using the relative entropy (Garcia *et al.*, 2002) relatively high sonority values for the silence periods and for the fricative segments were obtained (due to the quasi-periodicity of the noise).

The normalized Euclidean distance was chosen because it has the same characteristics of the relative entropy, while eliminating its drawbacks: giving low sonority values for the silence periods and better behaviour for the fricative segments. A problem with this metric was the fact that it returned high differences between the sonority of voiced regions and that of the unvoiced regions (by several orders of magnitude) which posed problems when computing the inverse of the function. This issue was solved by introducing a variable normalizing factor β, that reduces the distances between the values for the voiced and unvoiced regions while still keeping a significant difference between them.

At the beginning of the second step the envelope of the sonority is computed in order to find the sonority maxima. The envelope is computed by low-pass filtering the sonority function, thus obtaining only the long-term variations of the signal. Because the normalized Euclidean distance has a much smoother form than the relative entropy, it needs also a less

sophisticated method for computing the envelope. The syllable boundaries will be placed in accordance with the SSP. As it states that the peaks in the sonority function correspond to the syllable nuclei, by finding the maxima of the function, we will be finding the syllable nuclei. This is done first by imposing a minimum threshold on the sonority maxima and then searching for all local maxima in the signal. Having found the syllable nuclei and knowing that the syllable boundaries correspond to the minima in the sonority function, the next step consists in finding the minima between each two consecutive maxima.

The post-processing step tries to correct some of the errors that might appear in the segmentation process. In a first stage, a voice activity detection procedure is used to determine the beginning and the end of the speech region and all syllables boundaries found outside this interval are eliminated. One of the most important errors is the existence of spurious maxima close to the syllable nuclei, due to the semi-vowels, nasals or liquids that are in the vicinity of the vowel. Because the minima between two such maxima has very high values, these types of errors can be corrected by comparing each minima with their neighbouring maxima and eliminating the lower maximum in case of an insertion.

Another type of error might appear due to segments violating the SSP, in which a more sonorous segment is found further away from the nucleus than a less sonorous segment. In this case, our system tends to consider this segment as a unique syllable. But, due to the fact that these segments are quite short (usually under 75 ms), by setting a minimum threshold to the syllable length and assigning these isolated segments to one of the neighbouring syllables. For this value of the threshold the speech rate is not an issue as a speech rate of 14 syllables/second (corresponding to syllables 75 ms long) is difficult, if not impossible to be reached.

The erroneously found syllables corresponding to nasal segments are corrected by taking into consideration one of the most important acoustic characteristics of nasal consonants, i.e. the existence of a high intensity F1 around 300 Hz. In the case of boundaries with relatively high sonority and one of the syllables it confines short enough, a comparison between the F1 of this short segment and its neighbouring segment is done. If its mean F1 value is lower than the one of its neighbouring segment and also lower than 400 Hz the boundary will be deleted.

4. RESULTS

The corpora on which our system was tested are the Italian part of the SPEECON corpus (Siemund *et al.*, 2000) and the Switchboard corpus (Godfrey *et al.*, 1992). The Italian part of the SPEECON corpus contains numbers from 0 to 999,999 pronounced by male speakers. There are a total of 1906 recordings made by approximately 400 speakers. The Switchboard corpus instead is a corpus of English conversational speech recorded over the telephone line. It contains 2500 conversations collected from 500 American English speakers (both males and females).

The evaluation of the segmentation was performed using the algorithm presented in (Petek *et al.*, 1996). The algorithm defines for each of the manually annotated syllable boundaries a search region in which a corresponding automatically found syllable boundary will be searched for. The search interval spans from the middle of the interval between the

previous and the current syllable boundary to the middle of the interval between the current and the next syllable. If no automatic boundary is found in the search interval a deletion is considered. If a boundary is found, depending on the distance to the closest manual boundary, it is considered either a correct boundary or a substitution. If more automatic boundaries are found in the search interval, the closest one is considered the correct one and all the others are considered insertions.

In Table 1 we present a summary of the accuracies obtained using several algorithms for automatic syllable segmentation, while in Table 2 we show a comparison of the errors obtained between our system and the approaches presented in (Nagarajan *et al.*, 2003) and (Petrillo & Cutugno, 2003).

In each cell of Table 2 we have the substitutions, insertions and deletions that occurred during the segmentation process. Substitutions were considered if the distance between the found boundary and the manual one exceeds 40 ms, unless stated otherwise.

Corpus→ ↓Approach	Switchboard [%]	SPEECON [%]	Other [%]
Our algorithm	54.11	77.70	-
Mayora-Ibarra & Curatelli	-	-	SPK-IRST (Italian digits) 95[1]
Kawai & van Santen	-	-	Read news 62[2]
Nagarajan *et al.*	74.84	-	-
Petrillo & Cutugno	57.25	82.43	-

Table 1: Accuracies obtained using different segmentation algorithms

Corpus→ ↓Approach	Switchboard sub/ins/del [%]	SPEECON sub/ins/del [%]
Our algorithm	15.33 / 14.97 / 15.58	10.31 / 7.84 / 4.15
Nagarajan *et al.*	12.79 / 5.25 / 7.1	-
Petrillo & Cutugno	13.67 / 8.88 / 20.2	8.74 / 4.29 / 4.55

Table 2: Errors obtained using different segmentation algorithms

1 The error interval was set at 15 ms
2 For syllable nuclei

Although our approach gives accuracies values far from the state of the art system presented in Nagarajan *et al.* (2003), we obtain closer values to the systems using less complex speech processing techniques. The accuracies, both on the SPEECON corpus as on the Switchboard corpus, close to the modified system (Petrillo & Cutugno, 2003) as well as a much better accuracy (of the syllable boundaries) with respect to the Kawai & van Santen (2002) approach are an encouraging result.

5. CONCLUSIONS AND FUTURE WORK

A syllable segmentation algorithm based on the sonority of the speech signal was presented. The algorithm, which is based entirely on the signal, without any linguistic or phonetic knowledge, uses the Sonority Sequencing Principle for finding the syllable boundaries, corresponding to the minima in the sonority. The results obtained are encouraging, having obtained better accuracies than previous systems based on the sonority and accuracies similar to those of systems based on the energy of the signal.

In order to increase the segmentation accuracy, several other features could be used together with the sonority function. An example of such features are the acoustic properties of the signal that help us characterizing the manner of articulation of the speech segments included in the utterance. These can be in particular useful in the case of SSP violation – for example the presence of /s/ before /t/ in the onset of the syllable. By finding an isolated strident segment between two syllables and the second syllable beginning with a very low sonority segment (a stop consonant) we could consider it as a onset /s/ and assign it to the second syllable.

Another alternative would be the combination of our system with the one presented in Petrillo & Cutugno (2003). Initial tests on the errors that the two systems do showed that most of the errors are quite complementary and a combination of the two systems will increase the segmentation accuracy.

ACKNOWLEDGEMENTS

This work was supported by the EU FP6 Marie Curie Research Training Network 'Sound to Sense'.

6. REFERENCES

Cassandro, M., Collet, P., Duarte, D., Galves, A. & Garcia, J. (2002), *An universal linear relation among acoustic correlates of rhythm*, Retrieved from http://www.ime.usp.br/~tycho/participants/a_galves/linearity.pdf.

Clements, G. (1990), The role of the sonority cycle in core syllabification, in *Papers in laboratory phonology 1: between the grammar and physics of speech* (J. Kingston & M. Beckman, editors), Cambridge: Cambridge University Press, 283-333.

Galves, A., Garcia, J., Duarte, D. & Galves, C. (2002), Sonority as a basis for rhythmic class discrimination, in *Proceedings of the Speech Prosody 2002*, Aix en Provence, France, 323-326.

Garcia, E., Gut, U.B. & Galves, A. (2002), Vocale – a semi-automatic annotation tool for prosodic research, in *Proceedings of the Speech Prosody 2002*, Aix en Provence, France, 327-330.

Godfrey, J.J., Holliman, E.C. & McDaniel, J. (1992), SWITCHBOARD: Telephone speech corpus for research and development, in *Proceedings of IEEE ICASSP 1992*, 517-520.

Harris, J. (2006), The phonology of being understood: further arguments against sonority, *Lingua*, 116, 1483-1494.

Jespersen, O. (1904), *Lehrbuch der Phonetik*, Leipzig: B.G. Teubner.

Kawai, G. & van Santen, J. (2002), Automatic detection of syllabic nuclei using acoustic measures, in *2002 IEEE Workshop on Speech Synthesis*, Santa Monica, California, 39-42.

Ladefoged, P. (1993), *A Course in Phonetics*, 3rd edition (International Edition), Orlando: Harcourt Brace & Company.

Mayora-Ibarra, O. & Curatelli, F. (2002), Time-Domain Segmentation and Labelling of Speech with Fuzzy-Logic Post-Correction Rules, in *Proceedings of the Second Mexican International Conference on Artificial intelligence: Advances in Artificial intelligence*, 1-14.

Nagarajan, T., Murthy, H.A. & Hegde, R. M. (2003), Segmentation of speech into syllable-like units, in *EUROSPEECH-2003*, Geneva, Switzerland, 2893-2896.

Petek, B., Andersen, O. & Dalsgaard, P. (1996), On the robust automatic segmentation of spontaneous speech, in *ICSLP-1996*, Philadelphia, USA, 913-916.

Petrillo, M. & Cutugno, F. (2003), A syllable segmentation algorithm for English and Italian, in *EUROSPEECH-2003*, Geneva, Switzerland, 2913-2916.

Siemund, R., Höge, H., Kunzmann, S. & Marasek, K. (2000), *SPEECON* - Speech Data for Consumer Devices, in *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, 2000, vol. 2, 883-886.

Sievers, E. (1881), *Grundzüge der Phonetik,* Leipzig: Breitkopf und Härtel.

STATICO VS. DINAMICO.
UN POSSIBILE RUOLO DELLA SILLABA
NEL RICONOSCIMENTO AUTOMATICO DEL PARLATO

Serena Soldo, Bogdan Ludusan
Università degli studi di Napoli "Federico II"
soldo@na.infn.it, ludusan@na.infn.it

1. SOMMARIO

Il presente lavoro si pone come obiettivo quello di esplorare possibili tecniche di rappresentazione delle sillabe. La proposta è quella di utilizzare le caratteristiche statiche del segnale che rappresenta una sillaba. I parametri per questo tipo di rappresentazione sono stati estratti secondo due tecniche diverse: usando un numero variabile di parametri oppure un numero fisso. Per scegliere il tipo di rappresentazione migliore è stato addestrato un classificatore SVM. Le prestazioni migliori sono state ottenute utilizzando 15 frames per sillaba, ciascuno rappresentato con 13 parametri MFCC, raggiungendo un'*accuracy* pari a 88,25%.

2. INTRODUZIONE

Il continuum fonico su cui un sistema automatico di riconoscimento deve lavorare viene normalmente segmentato in piccole porzioni sulle quali algoritmi basati su tecniche statistiche operano sia per l'identificazione dell'informazione linguistica in essi contenuta, sia per ricostruire a posteriori il contenuto complessivo dell'enunciato contenuto nel segnale acustico. Mentre tradizionalmente fino a pochi anni fa le dimensioni della porzione minima di analisi si aggiravano intorno a dimensioni che linguisticamente potremmo definire subfoniche, sempre più spesso, ormai, i sistemi di riconoscimento del parlato fanno uso di analisi di segmenti di parlato superiori ai 150-200 ms. Questa tendenza indica l'uso di parametri soprasegmentali oltre che segmentali. Fra i parametri per la descrizione di segmenti lunghi che è possibile usare si incontrano quelli legati a proprietà ritmiche del parlato. Recentemente sono stati portati avanti lavori per dimostrare che tali parametri possono essere estratti automaticamente con algoritmi indipendenti dalla lingua (Tamburini & Caini, 2005; Petrillo, 2000; Ludusan & Soldo, 2009).

Sebbene la definizione 'classica' di sillaba (ma i linguisti sanno bene quanto trovare una definizione condivisa da tutti sia difficile) solitamente utilizzata in letteratura tende a mettere in evidenza le caratteristiche dinamiche del segnale vocale come ad esempio la coarticolazione, in questo lavoro si è cercato di proporre una ipotesi alternativa. L'idea è quella di vedere la sillaba come una rappresentazione statica di un pezzo di parlato, una sorta di istantanea che contenga in sé unitariamente informazione che solitamente si ritiene di tipo tempo-variabile, che si estende su un determinato intervallo di tempo. Alla luce di questo tipo di rappresentazione, la variabile indipendente rispetto alla quale i fenomeni che osserviamo evolvono e sulla quale possiamo basare un sistema di riconoscimento del parlato non risulterà più essere il tempo, ma la sequenza di unità sillabiche. Supponendo di essere in grado di individuare con precisione gli estremi dell'intervallo su cui si estende ciascuna sillaba, si può dunque pensare di 'fotografarla' estraendone le caratteristiche nei punti salienti.

È da osservare che questo genere di rappresentazione della sillaba è completamente originale e mai proposto in letteratura. Lo scopo di questo lavoro è proprio quello di capire se si tratta di una tecnica in grado di fornire buoni risultati e, eventualmente, di evidenziarne i punti deboli.

3. STRUMENTI

3.1 Support Vector Machine e LIBSVM

Tra i classificatori supervisionati più noti in letteratura troviamo le cosiddette *Support Vector Machines* (SVM) (Boser *et al.*, 1992; Cortes & Vapnik, 1995; Vapnik, 1995). In genere l'uso più comune, e per il quale le SVM risultano particolarmente adatte, è la classificazione di oggetti appartenenti a due sole classi ma esse possono essere facilmente estese anche al caso di più classi. Inoltre, negli ultimi anni sono stati tentati approcci nell'uso delle SVM proprio per la classificazione del parlato (in particolare Ganapathiraju, 2002). Le SVM fanno parte della famiglia di classificatori a 'massimo margine', infatti hanno come scopo quello di individuare una superficie di decisione che sia il più lontana possibile da ciascun punto dell'insieme dei dati. Tale distanza rappresenta il 'margine' del classificatore. Il nome *Support Vector Machines* deriva dal ruolo fondamentale di un particolare insieme di dati chiamati 'vettori di supporto'. Questi vettori sono, in realtà, gli unici ad avere peso nella scelta della superficie di decisione. In figura 1 è mostrato un esempio di superficie di decisione (a sinistra) e la superficie di decisione ottima (a destra) con i corrispondenti vettori di supporto. Osserviamo che le SVM sono l'unica famiglia di classificatori che può garantire determinate prestazioni di generalizzazione. Infatti la teoria di Vapnik (Vapnik, 1995) dimostra che la soluzione a massimo margine è anche la soluzione a massima generalizzazione.

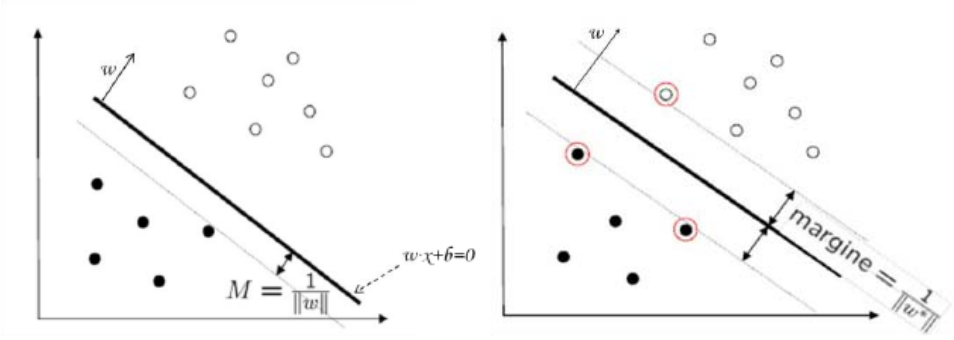


Figura 1: Esempio di un generico iperpiano di decisione (a sinistra) e l'iperpiano di decisione ottimo (a destra) per un problema di classificazione basato su due classi. Gli elementi cerchiati rappresentano i vettori di supporto

Formalmente possiamo definire il problema di classificare tramite SVM nel seguente modo: sia $S = (x_1; y_1), (x_2; y_2), \ldots(x_n; y_n)$ un training-set di punti tale che $x_i \in \Re^m$ e

$y_i \in \{-1; +1\}$ e sia $D(x) := \omega \cdot x + b = 0$ l'equazione per un generico iperpiano, si vuole risolvere il seguente problema di ottimizzazione:

$$\begin{aligned} &\min_{\omega,b} && \tfrac{1}{2}|\omega|^2 \\ &\text{s. a} && \\ &&& y_i(\omega \cdot x_i + b) \geq 1, \quad \forall i = 1 \ldots n \end{aligned} \tag{1}$$

La soluzione, ovvero il margine massimo, sarà data da $M^* = 1/|\omega^*|$. Tale soluzione può essere individuata con uno qualsiasi degli algoritmi ideati per i problemi di ottimizzazione con funzione obiettivo quadratica e vincoli lineari.

Nel caso in cui il training-set non risulti linearmente separabile, è possibile proiettarlo in un spazio di dimensione maggiore in cui aumenta la possibilità che i punti siano linearmente separabili (Cover, 1965). Le funzioni di Kernel hanno lo scopo simulare la proiezione dei punti in un nuovo spazio nel caso essi non siano linearmente separabili. Le funzioni di Kernel devono essere funzioni continue, simmetriche e definite positive. Tra le più usate funzioni di Kernel ci sono le funzioni a base radiale e le funzioni polinomiali.

Per la costruzione di un classificatore sillabico tramite SVM è stata usata una libreria specifica, LIBSVM (Chang & Lin, 2001), che mette a disposizione funzioni per il *training* del classificatore, lo *scaling* dei parametri, e la classificazione di nuovi oggetti.

La libreria è molto flessibile e permette l'uso di diversi kernel e la personalizzazione dei relativi parametri. Per quanto riguarda la classificazione 'multi-class' (ovvero con più di due classi), la libreria implementa la tecnica 'uno-contro-uno' poiché gli esperimenti degli autori (Hsu *et al.*, 2003) hanno mostrato che nonostante il numero di classificatori binari utilizzato sia maggiore, le prestazioni in termini di tempo e di risultati ottenuti sono decisamente migliori. Come vedremo nel paragrafo 4, dai test effettuati è emerso che, tra i kernel messi a disposizione dalla libreria, il kernel *Radial Basis Function (RBF)* è il più efficace per la classificazione del nostro spazio di dati. La caratteristica di questo kernel è quella di saper modellare aree di decisione chiuse. Questa proprietà è utile quando l'insieme dei dati si distribuisce nello spazio in modo che una classe sia completamente incapsulata in un'altra. Evidentemente la grossa somiglianza tra alcune coppie di sillabe (ad esempio 'di' e 'dje' oppure 'tre' e 'tren') rende fondamentale per la classificazione l'uso di un kernel come quello RBF.

3.2 Il corpus

Il corpus adottato in questo lavoro è una parte del corpus SPEECON (Siemund *et al.*, 2000); in questo corpus sono presenti 18 differenti lingue. La parte del corpus utilizzata riguarda i numeri in lingua italiana tra 0 e 999,999 pronunciati da soli speaker maschi; ogni registrazione audio ha una frequenza di 16 KHz e contiene l'enunciazione di un qualsiasi numero tra 0 e un milione (escluso). Le registrazioni audio sono 1906, pronunciate da circa 400 speaker differenti, i quali hanno registrato in media circa 5 file a testa. In base alla divisione sillabica fonologica (Cutugno *et al.*, 2001), gli enunciati contenuti nei file del corpus sono stati suddivisi in 8631 sillabe distinte che vanno a coprire l'intero insieme di 42 classi di sillabe presenti nel corpus. La tabella 1 mostra l'elenco delle sillabe e le relative occorrenze all'interno del corpus. Osserviamo che la sillaba 'due' costituisce un caso

particolare. Secondo Canepari "la distinzione tradizionale fra dittongo e iato è puramente teorica" (Canepari, 1999: 143). Se è vero che per [due] vs. [dwe] in forma isolata possono sussistere dei dubbi, è indubbio lo spostamento di accento in tutti i casi in cui il numero inizia con 'due' e segue: in questi casi 'due' cliticizza ed è sempre una sillaba.

Sillaba	**# occ.**	**Sillaba**	**# occ.**	**Sillaba**	**# occ.**
di	361	no	462	to	614
dje	67	o	177	tre	259
do	57	ran	62	tren	64
due	233	ro	119	ttan	111
dze	121	se	391	tte	283
sei	220	ssan	87	tto	293
kwa	360	sse	62	ttor	52
kwan	86	tʃa	114	ttro	246
kwe	215	tʃen	581	tu	53
kwin	50	tʃi	314	u	123
la	300	tʃin	300	un	57
lle	50	tʃo	55	van	57
mi	356	ta	367	ve	267
nno	52	ti	54	ven	61

Tabella 1: Elenco delle sillabe e relativo numero di occorrenze all'interno del *corpus*

4. METODO

4.1 Primo approccio: la sillaba come un volto

Il primo passo del nostro lavoro è consistito nella trasformazione di ogni porzione di segnale corrispondente ad una sillaba in un set di parametri (d'ora in poi *features*) da fornire in ingresso ad un sistema di riconoscimento.

Come è noto, una sillaba è costituita da almeno una vocale (che ne costituisce il nucleo) e può al massimo essere formata da tre parti, il nucleo vocalico testé definito, la testa e la coda. Per rappresentare ciascuna sillaba abbiamo quindi scelto di concentrare l'estrazione delle *features* solo sul centro di ciascuna delle tre parti. In particolare sono stati estratti tre vettori di parametri per la testa, tre per il nucleo e tre per la coda. Si è scelto di utilizzare per la rappresentazione i 13 coefficienti MFCC (*Mel Frequency Cepstral Coefficients*). Ciascuna sillaba, alla luce di queste scelte, risulta essere rappresentata da una matrice di dimensioni 9 x 13.

L'idea di rappresentare la sillaba in questo modo è nata dall'incontro con tecniche simili utilizzate nell'ambito del riconoscimento dei volti (ad esempio Samaria & Fallside, 1993). I tratti somatici si presentano sempre nello stesso ordine, indipendentemente dall'angolazione in cui si presenta il volto. Più precisamente, un viso può sempre essere diviso in 5

fasce orizzontali che individuano: Fronte, Occhi, Naso, Bocca, Mento. La nostra rappresentazione segue questo stesso principio spezzando ciascuna sillaba nei suoi 'tratti somatici' (testa, nucleo e coda) ed estraendo informazioni per ciascun segmento. La figura 2 mostra un esempio di tale suddivisione per un volto e per una sillaba.

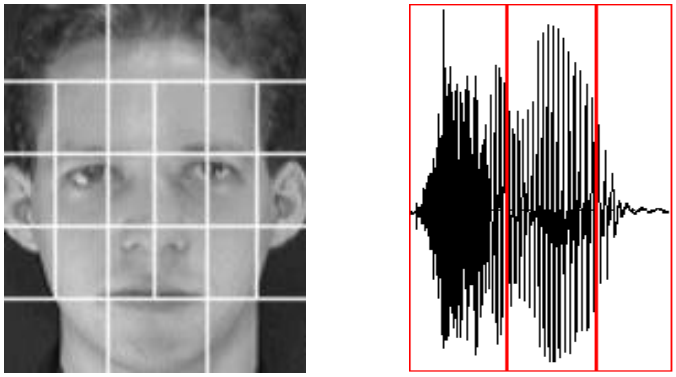


Figura 2: Esempio di segmentazione di un volto (a sinistra)
e di una sillaba (a destra) nei rispettivi 'tratti somatici'

Scelto questo tipo di rappresentazione, abbiamo addestrato una SVM multiclasse sulla base di un *training set* estratto dal corpus. In particolare l'SVM è stato addestrato sulle 42 classi di sillabe fonologiche presenti nel corpus e su una ulteriore classe che comprende il silenzio e le eventuali zone rumorose. L'efficacia della rappresentazione è stata poi valutata sul test set. Le prestazioni ottenute dalla classificazione delle sillabe rappresentate con questo primo approccio è pari all'85% circa (i risultati in dettaglio sono riportati nel § 4).

4.2 Evoluzioni

Le prestazioni ottenute dal classificatore utilizzando il metodo descritto al paragrafo precedente sono risultate incoraggianti anche se non particolarmente elevate a causa del fatto che la scelta di soli nove *frames* per la rappresentazione di qualsiasi sillaba non è forse la più adatta. Quindi una delle ipotesi prese in considerazione è stata quella di cambiare il numero di *frames* utilizzati. Le possibili tecniche da utilizzare a questo scopo sono due: il numero di *frames* considerati per ciascun segmento può variare al variare della dimensione del segmento stesso; oppure il numero di *frames* può essere fissato a priori ma con un valore più alto, scegliendo una maggiore o minore sovrapposizione delle finestre di avanzamento durante l'estrazione delle *features* in modo da adattarsi a ciascun segmento. Nella realizzazione della prima tecnica la sovrapposizione tra le finestre durante l'estrazione delle *features* è stata mantenuta fissa e quindi per ogni segmento sono stati estratti un numero di vettori variabile in base alla lunghezza del segmento stesso. Per questa tecnica sono state fatte delle prove considerando finestre di 128, 256 e 512 campioni (corrispondenti a circa 8, 16 e 32 msec), ma entrambe hanno prodotto risultati molto scarsi; per tale motivo questo sistema è stato accantonato in favore della seconda tecnica, molto più promettente. La seconda tecnica mira ad ottenere un numero di *frames* fisso da ciascun segmento; questo è stato ottenuto fissando l'ampiezza della finestra (nel nostro caso 256 campioni, corrispondenti a 16 msec) e variando l'ampiezza della sovrapposizione opportunamente.

5. RISULTATI

Le due tecniche di rappresentazione delle sillabe sono state testate tramite un classificatore SVM opportunamente addestrato. Le tabelle 2 e 3 riportano nel dettaglio i valori di *accuracy* ottenuti per ciascuna variante.

Shift tra i *frames* (ms)	Kernel Lineare [%]	Kernel Polinomiale [%]	Kernel RBF [%]
8	78,11	74,06	65,50
16	77,61	73,70	65,46
32	75,13	71,09	63,92

Tabella 2: Prestazione della classificazione per i diversi tipi di kernel utilizzando un numero variabile di frames per ogni sillaba

Numero di *frames* per sillaba	Kernel Lineare [%]	Kernel Polinomiale [%]	Kernel RBF [%]
15	86,53	85,88	88,25
17	85,85	85,63	88,10
19	85,95	85,60	88,21
21	85,99	85,81	88,75
23	85,81	85,56	88,18
25	86,10	85,99	88,32

Tabella 3: Prestazione della classificazione per i diversi tipi di kernel utilizzando un numero fisso di frames per ogni sillaba

Come si può notare, l'uso di un numero di *frames* variabile si è dimostrata una tecnica completamente inefficace mentre l'uso di un numero fisso di *frames* per sillaba è risultato decisamente migliore. In particolare, abbiamo fatto variare il numero di *frames* per ogni sillaba tra 15 e 25; le prestazioni ottenute all'aumentare del numero di *frames* non sono risultate significativamente migliori. Per tale motivo abbiamo fissato a 15 il numero di *frames* ideale per questo genere di rappresentazione.

Un'ulteriore analisi che si può fare sui risultati della classificazione è la valutazione degli N-best. Un classificatore SVM, nel tentativo di predire la classe da attribuire ad un elemento, individua la probabilità di appartenenza dell'elemento stesso ad ogni possibile classe: la classe con probabilità più alta è quella restituita in output. Facendo in modo che la SVM restituisca in output anche tutte le coppie <classe, probabilità> per ogni elemento da classificare e per ogni classe, è possibile valutare in che posizione di questa 'classifica' si trova la classe corretta di appartenenza di ciascun elemento. La tabella 4 riassume i risultati emersi da questa analisi. In particolare si osserva che nel 96% dei casi la classe giusta rientra tra le prime 3 più probabili e nel 99% dei casi essa è nelle prime 10.

N-best	*Accuracy* [%]
1	88.25
3	96.67
5	98.14
10	99.18
30	99.96

Tabella 4: Percentuali di *accuracy* nella valutazione degli N-best

6. DISCUSSIONE

I risultati ottenuti incoraggiano la costruzione di un algoritmo di decodifica che combini le informazioni fornite dal classificatore con quelle 'top down' provenienti dal dizionario usato per il riconoscimento fornendo in output la sequenza di sillabe che ha più probabilità di essere contenuta nel segnale da riconoscere.

In conclusione questo lavoro ha indagato nuove tecniche di estrazione delle features per la rappresentazione delle sillabe. Abbiamo mostrato come un approccio teso a estrarre informazioni sulle caratteristiche statiche del segnale, piuttosto che quelle dinamiche, può fornire buoni risultati.

7. BIBLIOGRAFIA

Boser, B.E., Guyon, I.M., & Vapnik, V.N. (1992), A training algorithm for optimal margin classifiers, in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, Pittsburgh, Pennsylvania, July 27-29, 1992, 144-152.

Canepari, L. (1999), *Manuale di Pronuncia Italiana*, Bologna: Zanichelli.

Chang, C.C. & Lin, C.J. (2001), LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cortes, C., & Vapnik, V. (1995), Support vector networks, *Machine Learning*, 273-297.

Cover, T.M. (1965), Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Transactions on Electronic Computers,* 14(3), 326-334.

Cutugno, F., Passaro, G. & Petrillo, M. (2001), Sillabificazione fonologica e sillabificazione fonetica, in *Dati empirici e teorie linguistiche*, Atti del XXXIII Congresso della Società di Linguistica Italiana, Napoli, 28-30 ottobre 1999 (F. Albano Leoni, R. Sornicola, E. Stenta Krosbakken, C. Stromboli, editors), Roma: Bulzoni, 205-232.

Ganapathiraju, A. (2002), *Support Vector Machines for Speech Recognition,* PhD thesis, Faculty of Mississippi State University.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003), *A practical guide to support vector classification,* Technical report, Taipei, Taiwan: National Taiwan University.

Ludusan, B. & Soldo, S. (2009), Sonority based syllable segmentation, in *La dimensione temporale del parlato* (S. Schmid, M. Schwarzenbach & D. Studer, editors), Atti del 5° Convegno Nazionale dell' Associazione Italiana di Scienze della Voce, Zurigo, Svizzera, 4-6 febbraio 2009, 699-706 (in questo volume).

Petrillo, M. (2000), *Algoritmi per la divisione del segnale verbale in unità sillabiche*, Tesi di Laurea presso l'Università degli Studi Di Napoli "Federico II".

Samaria, F. & Fallside, F. (1993), Face Identification and Feature Extraction Using Hidden Markov Models, in *Image Processing: Theory and Applications*, 1, (G. Vernazza, editor), Amsterdam: Elsevier, 295-298.

Siemund, R., Höge, H., Kunzmann, S., & Marasek, K. (2000), Speecon - Speech data for consumer devices, in *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, May 31-June 2, 2000, 883-886.

Tamburini, F. & Caini, C. (2005), An Automatic System for Detecting Prosodic Prominence in American English Continuous Speech, *International Journal of Speech Technology 2005*, 8, 33 – 44.

Vapnik, V.N. (1995), *The Nature of Statistical Learning Theory*, New York: Springer.